

Choosing a penalty for model selection in heteroscedastic regression

Sylvain Arlot *

CNRS ; Willow Project-Team

Laboratoire d'Informatique de l'Ecole Normale Supérieure

(CNRS/ENS/INRIA UMR 8548)

INRIA - 23 avenue d'Italie - CS 81321

75214 PARIS Cedex 13 - France

`sylvain.arlot@ens.fr`

June 3, 2010

1 Introduction

Penalization is a classical approach to model selection. In short, penalization chooses the model minimizing the sum of the empirical risk (how well the model fits data) and of some measure of complexity of the model (called penalty); see FPE [1], AIC [2], Mallows' C_p or C_L [22]. A huge amount of literature exists about penalties proportional to the dimension of the model in regression, showing under various assumption sets that dimensionality-based penalties like C_p are asymptotically optimal [26, 21, 24], and satisfy non-asymptotic oracle inequalities [12, 10, 11, 13]. Nevertheless, all these results assume data are homoscedastic, that is, the noise-level does not depend on the position in the feature space, an assumption often questionable in practice. Furthermore, C_p is empirically known to fail with heteroscedastic data, as showed for instance by simulation studies in [6, 8].

In this paper, it is assumed that data *can* be heteroscedastic, but not necessary with certainty. Several estimators adapting to heteroscedasticity have been built thanks to model selection (see [19] and references therein), but always assuming the model collection has a particular form. Up to the best of our knowledge, only cross-validation or resampling-based procedures are built for solving a general model selection problem when data are heteroscedastic. This fact was recently confirmed, since resampling and V -fold penalties satisfy oracle inequalities for regressogram selection when data are heteroscedastic [6, 5]. Nevertheless, adapting to heteroscedasticity with resampling usually implies a significant increase of the computational complexity.

The main goal of the paper is to understand whether the additional computational cost of resampling can be avoided, when, and at which price in terms of statistical performance. Let us emphasize that determining from data only whether the noise-level is constant is a difficult question, since variations of the noise can easily be interpreted as variations of the smoothness of the signal, and conversely. Therefore, the problem of choosing an appropriate penalty—in particular, between dimensionality-based and resampling-based penalties—must be solved unless

*<http://www.di.ens.fr/~arlot/>

homoscedasticity of data is not questionable at all. The answer clearly depends at least on what is known about variations of the noise-level, and on the computational power available.

The framework of the paper is least-squares regression with a random design, see Section 2. We assume the goal of model selection is efficiency, that is, selecting a least-squares estimator with minimal quadratic risk, without assuming the regression function belongs to any of the models. Since we deal with a non-asymptotic framework, where the collection of models is allowed to grow with the sample size, a model selection procedure is said to be optimal (or efficient) when it satisfies an oracle inequality with leading constant (asymptotically) one. A classical approach to design optimal procedures is the unbiased risk estimation principle, recalled in Section 3.

The main results of the paper are stated in Section 4. First, all dimensionality-based penalties are proved to be suboptimal—that is, the risk of the selected estimator is larger than the risk of the oracle multiplied by a factor $\mathbb{C}_1 > 1$ —as soon as data are heteroscedastic, for selecting among regressogram estimators (Theorem 2). Note that the restriction to regressograms is merely technical, and we expect a similar result holds for general heteroscedastic model selection problems. Compared to the oracle inequality satisfied by resampling-based penalties in the same framework (Theorem 1, recalled in Section 3), Theorem 2 shows what is lost when using dimensionality-based penalties with heteroscedastic data: at least a constant factor $\mathbb{C}_1 > 1$.

Second, Proposition 2 shows that a well-calibrated penalty proportional to the dimension of the models does not loose more than a constant factor $\mathbb{C}_2 > \mathbb{C}_1$ compared to the oracle. Nevertheless, C_p strongly overfits for some heteroscedastic model selection problems, hence loosing a factor tending to infinity with the sample size compared to the oracle (Proposition 3). Therefore, a proper calibration of dimensionality-based penalties is absolutely required when heteroscedasticity is suspected.

These theoretical results are completed by a simulation experiment (Section 5), showing a slightly more complex finite-sample behaviour. In particular, when the signal-to-noise ratio is rather small, improving a well-calibrated dimensionality-based penalty requires a significant increase of the computational complexity.

Finally, from the results of Sections 4 and 5, Section 6 tries to answer the central question of the paper: How to choose the penalty for a given model selection problem, taking into account prior knowledge on the noise-level and the computational power available?

All the proofs are made in Section 7.

2 Framework

In this section, we describe the least-squares regression framework, model selection and the penalization approach. Then, typical examples of collections of models and heteroscedastic data are introduced.

2.1 Least-squares regression

Suppose we observe some data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$, independent with common distribution P , where the feature space \mathcal{X} is typically a compact subset of \mathbb{R}^k . The goal is to predict Y given X , where $(X, Y) \sim P$ is a new data point independent of $(X_i, Y_i)_{1 \leq i \leq n}$. Denoting by s the regression function, that is $s(x) = \mathbb{E}[Y \mid X = x]$, we can write

$$Y_i = s(X_i) + \sigma(X_i)\varepsilon_i \tag{1}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise level and $(\varepsilon_i)_{1 \leq i \leq n}$ are i.i.d. centered noise terms; ε_i may depend on X_i , but has mean 0 and variance 1 conditionally on X_i .

The quality of a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$ is measured by the quadratic prediction loss

$$\mathbb{E}_{(X,Y) \sim P} [\gamma(t, (X, Y))] =: P\gamma(t) \quad \text{where} \quad \gamma(t, (x, y)) = (t(x) - y)^2$$

is the least-squares contrast. The minimizer of $P\gamma(t)$ over the set of all predictors, called Bayes predictor, is the regression function s . Therefore, the excess loss is defined as

$$\ell(s, t) := P\gamma(t) - P\gamma(s) = \mathbb{E}_{(X,Y) \sim P} (t(X) - s(X))^2 .$$

Given a particular set of predictors S_m (called a *model*), the best predictor over S_m is defined by

$$s_m := \operatorname{argmin}_{t \in S_m} \{P\gamma(t)\} .$$

The empirical counterpart of s_m is the well-known *empirical risk minimizer*, defined by

$$\hat{s}_m := \operatorname{argmin}_{t \in S_m} \{P_n\gamma(t)\}$$

(when it exists and is unique), where $P_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$ is the empirical distribution function; \hat{s}_m is also called least-squares estimator since γ is the least-squares contrast.

2.2 Model selection, penalization

Let us assume that a family of models $(S_m)_{m \in \mathcal{M}_n}$ is given, hence a family of empirical risk minimizers $(\hat{s}_m)_{m \in \mathcal{M}_n}$. The model selection problem consists in looking for some data-dependent $\hat{m} \in \mathcal{M}_n$ such that $\ell(s, \hat{s}_{\hat{m}})$ is as small as possible. For instance, it would be convenient to prove an oracle inequality of the form

$$\ell(s, \hat{s}_{\hat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\} + R_n \quad (2)$$

in expectation or with large probability, with leading constant C close to 1 and $R_n = \mathcal{O}(n^{-1})$.

This paper focuses more precisely on model selection procedures by penalization, which can be described as follows. Let $\text{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$ be some penalty function, possibly data-dependent, and define

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \{\text{crit}(m)\} \quad \text{with} \quad \text{crit}(m) := P_n\gamma(\hat{s}_m) + \text{pen}(m) . \quad (3)$$

The penalty $\text{pen}(m)$ can usually be interpreted as a measure of the size of S_m . Since the ideal criterion $\text{crit}(m)$ is the true prediction error $P\gamma(\hat{s}_m)$, the *ideal penalty* is

$$\text{pen}_{\text{id}}(m) := P\gamma(\hat{s}_m) - P_n\gamma(\hat{s}_m) .$$

This quantity is unknown because it depends on the true distribution P . A natural idea is to choose $\text{pen}(m)$ as close as possible to $\text{pen}_{\text{id}}(m)$ for every $m \in \mathcal{M}_n$. This idea leads to the well-known unbiased risk estimation principle, which is properly introduced in Section 3.1. For instance, when each model S_m is a finite dimensional vector space of dimension D_m and the noise-level is constant equal to σ , Mallows [22] proposed the C_p penalty defined by

$$\text{pen}_{\text{Cp}}(m) = \frac{2\sigma^2 D_m}{n} .$$

Penalties proportional to D_m , like C_p , are extensively studied in Section 4.

Among the numerous families of models that can be used, this paper mostly considers “histogram models”, where for every $m \in \mathcal{M}_n$, S_m is the set of piecewise constant functions w.r.t. some fixed partition Λ_m of \mathcal{X} . Note that the least-squares estimator \hat{s}_m on some histogram model S_m is also called a regressogram. Then, S_m is a vector space of dimension $D_m = \text{Card}(\Lambda_m)$ generated by the family $(\mathbf{1}_\lambda)_{\lambda \in \Lambda_m}$. Model selection among a family $(S_m)_{m \in \mathcal{M}_n}$ of histogram models amounts to select a partition of \mathcal{X} among $\{\Lambda_m\}_{m \in \mathcal{M}_n}$.

Three arguments motivate the choice of histogram models for this theoretical study. First, better intuitions can be obtained on the role of variations of the noise-level $\sigma(\cdot)$ over \mathcal{X} —or variations of the smoothness of s —because an histogram models is generated by a localized basis $(\mathbf{1}_\lambda)_{\lambda \in \Lambda_m}$. Second, histograms have good approximation properties when the regression function s is α -Hölderian with $\alpha \in (0, 1]$. Third, all important quantities for understanding the model selection problem can be precisely controlled and compared, see [5].

2.3 Examples of histogram model collections

Let us assume in this section for simplicity that $\mathcal{X} = [0, 1]$. We define in this section several collections of models $(S_m)_{m \in \mathcal{M}_n}$, always assuming that each S_m is the histogram model associated to some partition Λ_m of \mathcal{X} .

The most natural (and simple) collection of histogram models is the collection of *regular histograms* $(S_m)_{m \in \mathcal{M}_n^{(\text{reg})}}$ defined by

$$\forall m \in \mathcal{M}_n^{(\text{reg})} := \{1, \dots, M_n\}, \quad \Lambda_m = \left(\left[\frac{k-1}{m}, \frac{k}{m} \right) \right)_{1 \leq k \leq m},$$

where the maximal dimension $M_n \leq n$ usually grows with n slightly slower than n ; reasonable choices are $M_n = \lfloor n/2 \rfloor$, $M_n = \lfloor n/(\ln(n)) \rfloor$ or $M_n = \lfloor n/(\ln(n))^2 \rfloor$.

Model selection among the collection of regular histograms then amounts to selecting the number $D_m \in \{1, \dots, M_n\}$ of bins, or equivalently, selecting the bin size $1/D_m$ among $\{1/M_n, \dots, 1/2, 1\}$. The regular collection $\mathcal{M}_n^{(\text{reg})}$ is a good choice when the distribution of X is close to the uniform distribution on $[0, 1]$, the noise-level $\sigma(\cdot)$ is almost constant on $[0, 1]$, and the variations of s (measured by $|s'|$) are almost constant over \mathcal{X} .

Since we can seldom be sure these three assumptions are satisfied by real data, considering other collection of histograms models can be useful in general, in particular for adapting to possible heteroscedasticity of data, which is the main topic of the paper. The simplest case of collection of histogram models with variable bin size is the collection of *histograms with two bin sizes and split at $1/2$* , $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$, defined by

$$\mathcal{M}_n^{(\text{reg}, 1/2)} := \left\{ (D_1, D_2) \text{ s.t. } 1 \leq D_1, D_2 \leq \frac{M_n}{2} \right\} \cup \{1\},$$

where S_1 is the set of constant functions on \mathcal{X} and for every $m = (D_{m,1}, D_{m,2}) \in (\mathbb{N} \setminus \{0\})^2$,

$$\Lambda_m := \left\{ \left[\frac{k-1}{2D_{m,1}}, \frac{k}{2D_{m,1}} \right) \right\}_{1 \leq k \leq D_{m,1}} \cup \left\{ \left[\frac{D_{m,2} + k - 1}{2D_{m,2}}, \frac{D_{m,2} + k}{2D_{m,2}} \right) \right\}_{1 \leq k \leq D_{m,2}}.$$

Note that using a collection of models such as $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$ does not mean that data are *known* to be heteroscedastic; $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$ can also be useful when one only *suspects*

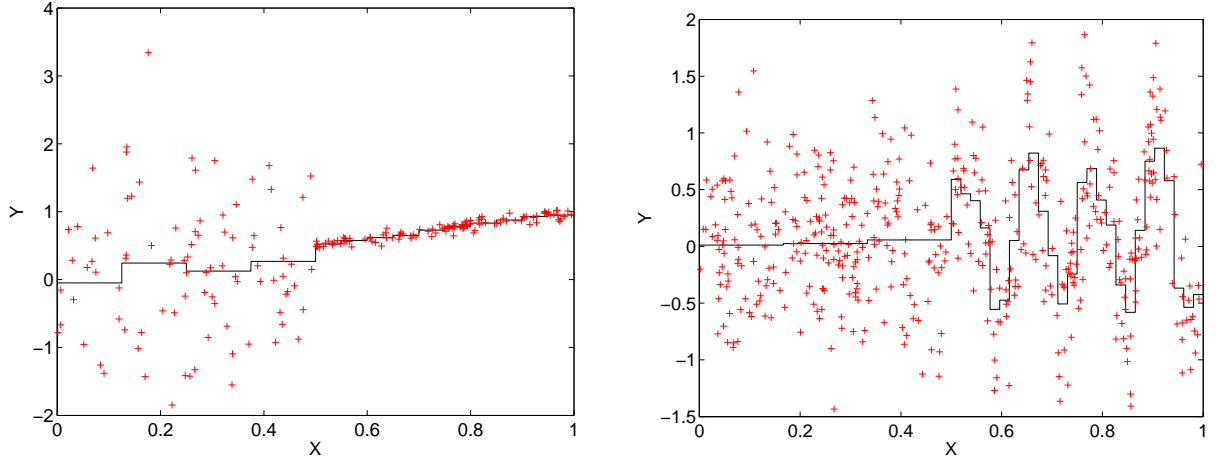


Figure 1: One data sample (red '+') and the corresponding oracle estimators (black '-') among the family $(\hat{s}_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$. Left: heteroscedastic data ($s(x) = x$; $\sigma(x) = 1$ if $x \leq 1/2$ and $\sigma(x) = 1/20$ otherwise) with sample size $n = 200$. Right: homoscedastic data ($s(x) = x/4$ if $x \leq 1/2$ and $s(x) = 1/8 + 2/3 \times \sin(16\pi x)$ otherwise; $\sigma(x) = 1/2$) with sample size $n = 500$.

that at least one quantity among σ , $|s'|$ and the density of X w.r.t. the Lebesgue measure (Leb) significantly varies over \mathcal{X} . Distinguishing between the three phenomena precisely is the purpose of model selection: Overfitting occurs when a large noise level is interpreted as large variations of s with low noise. The interest of using $\mathcal{M}_n^{(\text{reg}, 1/2)}$ is illustrated by Figure 1, where two data samples and the corresponding oracle estimators are plotted (left: heteroscedastic with s' constant; right: homoscedastic with s' variable).

Using only two different bin sizes, with a fixed split at $1/2$, obviously is not the only collection of histograms that may be used. Let us mention here a few examples of alternative histogram collections:

- the split can be put at any fixed position $t \in (0, 1)$ (possibly with different maximal number of bins $M_{n,1}$ and $M_{n,2}$ on each side of t), leading to the collection $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, t)}}$.
- the position of the split can be variable:

$$\mathcal{M}_n^{(\text{reg}, \text{var})} = \bigcup_{t \in \mathcal{T}_n} \mathcal{M}_n^{(\text{reg}, t)} \quad \text{where} \quad \mathcal{T}_n \subset (0, 1) \quad , \quad \text{for instance} \quad \mathcal{T}_n = \left\{ \frac{k}{\sqrt{n}} \text{ s.t. } 1 \leq k \leq \sqrt{n} - 1 \right\} .$$

- instead of a single split, one could consider collections with several splits (fixed or not), such that $\{1/3, 2/3\}$ or $\{1/4, 1/2, 3/4\}$ for instance.

Remark that $\text{Card}(\mathcal{M}_n^{(\text{reg}, 1/2)}) \leq M_n^2 \leq n^2$, and the cardinalities of all other collections are smaller than some power of n . Therefore, as explained in Section 3 below, penalization procedures using an estimator of $\text{pen}_{\text{id}}(m)$ for every $m \in \mathcal{M}_n$ as a penalty are relevant. This paper does not consider collections whose cardinalities grow faster than some power of n , such as the ones used for multiple change-point detection. Indeed, the model selection problem is of different nature for such collections, and requires the use of different penalties; see for instance [8] about this particular problem.

Most results of the paper are proved for model selection among $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$, which already captures most of the difficulty of model selection when data are heteroscedastic. The simplicity of $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$ may be a drawback for analyzing real data; in the present theoretical study, simplicity helps developing intuitions about the general case. Note that all results of the paper can be proved similarly when $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, t)}$ and $t \in (0, 1)$ is fixed; we conjecture these results can be extended to $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, \text{var})}$, at the price of additional technicalities in the proofs.

3 Unbiased risk estimation principle for heteroscedastic data

The unbiased risk estimation principle is among the most classical approaches for model selection [27]. Let us first summarize it in the general framework.

3.1 General framework

Assume that for every $m \in \mathcal{M}_n$, $\text{crit}(m, (X_i, Y_i)_{1 \leq i \leq n})$ estimates unbiasedly the risk $P\gamma(\hat{s}_m)$ of the estimator \hat{s}_m . Then, an oracle inequality like (2) with $C \approx 1$ should be satisfied by any minimizer \hat{m} of $\text{crit}(m, (X_i, Y_i)_{1 \leq i \leq n})$ over $m \in \mathcal{M}_n$. For instance, FPE [1], SURE [27] and cross-validation [3, 28, 18] are model selection procedures built upon the unbiased risk estimation principle.

When $\text{crit}(m, (X_i, Y_i)_{1 \leq i \leq n})$ is a penalized empirical criterion given by (3), the unbiased risk estimation principle can be rewritten as

$$\forall m \in \mathcal{M}_n, \quad \text{pen}(m) \approx \mathbb{E}[\text{pen}_{\text{id}}(m)] = \mathbb{E}[(P - P_n)\gamma(\hat{s}_m)] ,$$

which is also known as Akaike's heuristics or Mallows' heuristics. For instance, AIC [2], C_p or C_L [22] (see Section 4.1), covariance penalties [17] and resampling penalties [16, 6] are penalties built upon the unbiased risk estimation principle.

The unbiased risk estimation principle can lead to oracle inequalities with leading constant $C = 1 + o(1)$ when n tends to infinity, by proving that deviations of $P\gamma(\hat{s}_m)$ around its expectation are uniformly small with large probability. Such a result can be proved in various frameworks as soon as the number of models grows at most polynomially with n , that is, $\text{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$ for some $c_{\mathcal{M}}, \alpha_{\mathcal{M}} > 0$; see for instance [13, 9] and references therein for recent results in this direction in the regression framework.

3.2 Histogram models

Let S_m be the histogram model associated with a partition Λ_m of \mathcal{X} . Then, the concentration inequalities of Section 7.3 show that for most models, the ideal penalty is close to its expectation. Moreover, the expectation of the ideal penalty can be computed explicitly thanks to Proposition 4, first proved in a previous paper [5]:

$$\mathbb{E}[\text{pen}_{\text{id}}(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} (2 + \delta_{n, p_\lambda}) \left((\sigma_\lambda^r)^2 + (\sigma_\lambda^d)^2 \right) \quad (4)$$

where for every $\lambda \in \Lambda_m$,

$$\begin{aligned} (\sigma_\lambda^r)^2 &:= \mathbb{E} \left[(Y - s(X))^2 \mid X \in \lambda \right] = \mathbb{E} \left[(\sigma(X))^2 \mid X \in \lambda \right] & (\sigma_\lambda^d)^2 &:= \mathbb{E} \left[(s(X) - s_m(X))^2 \mid X \in \lambda \right] \\ p_\lambda &:= \mathbb{P}(X \in \lambda) & \text{and} & \quad \forall n \in \mathbb{N}, \forall p \in (0, 1], \quad |\delta_{n, p}| \leq \min \left\{ L_1, \frac{L_2}{(np)^{1/4}} \right\} \end{aligned}$$

for some absolute constants $L_1, L_2 > 0$.

When data are homoscedastic, (4) shows that if $D_m, \min_{\lambda \in \Lambda_m} \{np_\lambda\} \rightarrow \infty$ and if $\|s - s_m\|_\infty \rightarrow 0$,

$$\forall m \in \mathcal{M}_n, \quad \mathbb{E}[\text{pen}_{\text{id}}(m)] \approx \frac{2\sigma^2 D_m}{n} + \frac{2}{n} \sum_{\lambda \in \Lambda_m} \left(\sigma_\lambda^d\right)^2 \approx \frac{2\sigma^2 D_m}{n} = \text{pen}_{\text{Cp}}(m),$$

so that C_p should yield good model selection performances by the unbiased risk estimation principle. When the constant noise-level σ is unknown, pen_{Cp} can still be used by replacing σ^2 by some unbiased estimator of σ^2 ; see for instance [11] for a theoretical analysis of the performance of C_p with some classical estimator of σ^2 .

On the contrary, when data are heteroscedastic, (4) shows that applying the unbiased risk estimation principle requires to take into account the variations of σ over \mathcal{X} . Without prior information on $\sigma(\cdot)$, building a penalty for the general heteroscedastic framework is a challenging problem, for which resampling methods have been successful.

3.3 Resampling-based penalization

The resampling heuristics [15] provides a way of estimating the distribution of quantities of the form $F(P, P_n)$, by building randomly from P_n several “resamples” with empirical distribution P_n^W . Then, the distribution of $F(P_n, P_n^W)$ conditionally on P_n mimics the distribution of $F(P, P_n)$. We refer to [6] for more details and references on the resampling heuristics in the context of model selection. Since $\text{pen}_{\text{id}}(m) = F_m(P, P_n)$, the resampling heuristics can be used for estimating $\mathbb{E}[\text{pen}_{\text{id}}(m)]$ for every $m \in \mathcal{M}_n$. Depending on how resamples are built, we can obtain different kinds of resampling-based penalties, in particular the following three ones.

First, bootstrap penalties [16] are obtained with the classical bootstrap resampling scheme, where the resample is an n -sample i.i.d. with common distribution P_n . Second, general exchangeable resampling schemes can be used for defining the family of (exchangeable) resampling penalties [6, 20]. Third, V -fold penalties [5] are a computationally efficient alternative to bootstrap and other exchangeable resampling penalties; they follow from the resampling heuristics with a subsampling scheme inspired by V -fold cross-validation.

Let us define here V -fold penalties, which are of particular interest because of their smaller computational cost when V is small. Let $V \in \{2, \dots, n\}$ and $(B_j)_{1 \leq j \leq V}$ be a fixed partition of $\{1, \dots, n\}$ such that $\sup_j |\text{Card}(B_j) - n/V| < 1$. For every j , define

$$P_n^{(-j)} = \frac{1}{n - \text{Card}(B_j)} \sum_{i \notin B_j} \delta_{(X_i, Y_i)}$$

and $\forall m \in \mathcal{M}_n, \quad \hat{s}_m^{(-j)} \in \arg\min_{t \in S_m} \left\{ P_n^{(-j)} \gamma(t) \right\}.$

Then, the V -fold penalty is defined by

$$\text{pen}_{\text{VF}}(m) := \frac{V-1}{V} \sum_{j=1}^V \left(P_n - P_n^{(-j)} \right) \gamma \left(\hat{s}_m^{(-j)} \right). \quad (5)$$

In the least-squares regression framework, exchangeable resampling and V -fold penalties have been proved in [6, 5] to satisfy an oracle inequality of the form (2) with leading constant $C = C(n) \rightarrow 1$ when $n \rightarrow \infty$. In order to state precisely one of these results, let us introduce a set of assumptions, called **(AS)_{hist}**.

Assumption set $(\mathbf{AS})_{\text{hist}}$. For every $m \in \mathcal{M}_n$, S_m is the set of piecewise constants functions on some fixed partition Λ_m of \mathcal{X} , and \mathcal{M}_n satisfies:

(P1) Polynomial complexity of \mathcal{M}_n : $\text{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Richness of \mathcal{M}_n : $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} = \text{Card}(\Lambda_{m_0}) \in [\sqrt{n}, c_{\text{rich}} \sqrt{n}]$.

Moreover, data $(X_i, Y_i)_{1 \leq i \leq n}$ are i.i.d. and satisfy:

(Ab) Data are bounded: $\|Y_i\|_{\infty} \leq A < \infty$.

(An) Uniform lower-bound on the noise level: $\sigma(X_i) \geq \sigma_{\min} > 0$ a.s.

(Ap) The bias decreases like a power of D_m : constants $\beta_- \geq \beta_+ > 0$ and $C_b^+, C_b^- > 0$ exist such that

$$\forall m \in \mathcal{M}_n, \quad C_b^- D_m^{-\beta_-} \leq \ell(s, s_m) \leq C_b^+ D_m^{-\beta_+}.$$

(Ar $_{\ell}^{\mathbf{X}}$) Lower regularity of the partitions for $\mathcal{L}(X)$: $D_m \min_{\lambda \in \Lambda_m} \{\mathbb{P}(X \in \lambda)\} \geq c_{r,\ell}^{\mathbf{X}}$.

Remark 1. Assumption set $(\mathbf{AS})_{\text{hist}}$ is shown to be mild and discussed extensively in [6]; we do not report such a discussion here because it is beyond the scope of the paper. In particular, when s is non-constant, α -Hölderian for some $\alpha \in (0, 1]$ and X has a lower bounded density with respect to the Lebesgue measure on $\mathcal{X} = [0, 1]$, assumptions (P1), (P2), (Ap) and (Ar $_{\ell}^{\mathbf{X}}$) are satisfied by all the examples of model collections given in Section 2.3 (see in particular [5] for a proof of the lower bound in (Ap) for regular partitions, which applies to the examples of Section 2.3 since they are “piecewise regular”). Note also that *all the results* of the present paper relying on $(\mathbf{AS})_{\text{hist}}$ also hold under various alternative assumption sets. For instance, (Ab) and (An) can be relaxed, see [6] for details.

Theorem 1 (Theorem 2 in [5]). Assume that $(\mathbf{AS})_{\text{hist}}$ holds true. Then, for every $V \geq 2$, a constant $K_0(V)$ (depending only on V and on the constants appearing in $(\mathbf{AS})_{\text{hist}}$) and an event of probability at least $1 - K_0(V)n^{-2}$ exist on which, for every

$$\begin{aligned} \hat{m}_{\text{penVF}} &\in \operatorname{argmin}_{m \in \mathcal{M}_n} \{P_n \gamma(\hat{s}_m) + \text{pen}_{\text{VF}}(m)\}, \\ \ell(s, \hat{m}_{\text{penVF}}) &\leq \left(1 + (\ln n)^{-1/5}\right) \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\}. \end{aligned}$$

In particular, V -fold penalization is asymptotically optimal: when n tends to infinity, the excess loss of the estimator \hat{m}_{penVF} is equivalent to the excess loss of the oracle estimator \hat{s}_{m^*} , defined by $m^* \in \operatorname{argmin}_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\}$. A result similar to Theorem 1 has also been proved for exchangeable resampling penalties in [6], under the same assumption set $(\mathbf{AS})_{\text{hist}}$. In particular, Theorem 1 is still valid when $V = n$. Let us emphasize that *general unknown variations of the noise-level $\sigma(\cdot)$* are allowed in Theorem 1.

Theorem 1—as well as its equivalent for exchangeable resampling penalties—mostly follows from the unbiased risk estimation principle presented in Section 3.1: For every model $m \in \mathcal{M}_n$, $\mathbb{E}[\text{pen}_{\text{VF}}(m)]$ is close to $\mathbb{E}[\text{pen}_{\text{id}}(m)]$ whatever the variations of $\sigma(\cdot)$, and deviations of $\text{pen}_{\text{VF}}(m)$ around its expectation can be properly controlled. The oracle inequality follows, thanks to (P1).

The main drawback of exchangeable resampling penalties, and even V -fold penalties, is their computational cost. Indeed, computing these penalties requires to compute for every $m \in \mathcal{M}_n$ a least-squares estimator \hat{s}_m several times: V times for V -fold penalties, at least n times for

exchangeable resampling penalties. Therefore, except in particular problems for which \widehat{s}_m can be computed fastly, all resampling-based penalties can be untractable when n is too large, except maybe V -fold penalties with $V = 2$ or 3 . Note that (V -fold) cross-validation methods suffer from the same drawback, in addition to their bias which makes them suboptimal when V is small, see [5].

Furthermore, Theorem 1 could suggest that the performance of V -fold penalization does not depend on V , so that the best choice always is $V = 2$ which minimizes the computational cost. Although this asymptotically holds true at first order, quite a different picture holds when the signal-to-noise ratio is small, according to the simulation studies of [5] and of Section 5 below. Indeed, the amplitude of deviations of $\text{pen}_{\text{VF}}(m)$ around its expectation decreases with V , so that the statistical performance of V -fold penalties can be much better for large V than for $V = 2$.

Remark that one could also define *hold-out penalties* by

$$\forall m \in \mathcal{M}_n, \quad \text{pen}_{\text{HO}}(m) := \frac{\text{Card}(I)}{n - \text{Card}(I)} \left(P_n - P_n^{(I)} \right) \gamma \left(\widehat{s}_m^{(I)} \right) \quad \text{where } I \subset \{1, \dots, n\} \text{ is deterministic,}$$

$$P_n^{(I)} = \frac{1}{\text{Card}(I)} \sum_{i \in I} \delta_{(X_i, Y_i)} \quad \text{and} \quad \forall m \in \mathcal{M}_n, \quad \widehat{s}_m^{(I)} \in \underset{t \in S_m}{\text{argmin}} \left\{ P_n^{(I)} \gamma(t) \right\},$$

which only requires to compute once \widehat{s}_m for each $m \in \mathcal{M}_n$. The proof of Theorem 1 can then be extended to hold-out penalties provided that $\min \{ \text{Card}(I), n - \text{Card}(I) \}$ tends to infinity with n fastly enough, for instance when $\text{Card}(I) \approx n/2$. Nevertheless, hold-out penalties suffer from a larger variability than 2-fold penalties, which leads to quite poor statistical performances.

Therefore, when computational power is strongly limited and the signal-to-noise ratio is small, it may happen that none of the above resampling-based model selection procedures is satisfactory in terms of both computational cost and statistical performance. The purpose of the next two sections is to investigate whether the dimensionality of the models, which is freely available in general, can be used for building a computationally cheap model selection procedure with reasonably good statistical performance, in particular compared to V -fold penalties with V small.

4 Dimensionality-based model selection

Dimensionality as a vector space is the only information about the size of the models that is freely available in general. So, when some penalty must be proposed, functions of the dimensionality D_m of model S_m are the most natural (and classical) proposals. This section intends to measure the statistical performance of such procedures for least-squares regression with heteroscedastic data.

4.1 Examples

As previously mentioned in Section 2.2, C_p defined by $\text{pen}_{C_p}(m) = 2\sigma^2 D_m/n$ is the among most classical penalties for least-squares regression [22]. C_p belongs to the family of *linear penalties*, that is, of the form

$$\widehat{K} D_m,$$

where \widehat{K} can either depend on prior information on P (for instance, the value σ of the—constant—noise-level) or on the sample only. A popular choice is $\widehat{K} = 2\widehat{\sigma}^2/n$, where $\widehat{\sigma}^2$ is an

estimator of the variance of the noise, see Section 6 of [10] for instance. Birgé and Massart [13] recently proposed an alternative procedure for choosing \widehat{K} , based upon the “slope heuristics”.

Refined versions of C_p have been proposed, for instance in [13, 11, 25]—always assuming homoscedasticity. Most of them are of the form

$$\text{pen}(m) = \widehat{F}(D_m) \quad (6)$$

where \widehat{F} depends on n and σ^2 , or an estimator $\widehat{\sigma^2}$ of σ^2 when σ^2 is unknown. The rest of the section focuses on *dimensionality-based penalties*, that is, penalties of the form (6).

4.2 Characterization of dimensionality-based penalties

Let us define, for every $D \in \mathcal{D}_n = \{D_m \text{ s.t. } m \in \mathcal{M}_n\}$,

$$\mathcal{M}_{\text{dim}}(D) := \operatorname{argmin}_{m \in \mathcal{M}_n \text{ s.t. } D_m = D} \{P_n \gamma(\widehat{s}_m)\} \quad \text{and} \quad \mathcal{M}_{\text{dim}} := \bigcup_{D \in \mathcal{D}_n} \mathcal{M}_{\text{dim}}(D) .$$

The following lemma shows that any dimensionality-based penalization procedure actually selects $\widehat{m} \in \mathcal{M}_{\text{dim}}$.

Lemma 1. *For every function $F : \mathcal{M}_n \mapsto \mathbb{R}$ and any sample $(X_i, Y_i)_{1 \leq i \leq n}$,*

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \{P_n \gamma(\widehat{s}_m) + F(D_m)\} \subset \mathcal{M}_{\text{dim}} .$$

proof of Lemma 1. Let $\widehat{m}_F \in \operatorname{argmin}_{m \in \mathcal{M}_n} \{P_n \gamma(\widehat{s}_m) + F(D_m)\}$. Then, whatever $m \in \mathcal{M}_n$,

$$P_n \gamma(\widehat{s}_{\widehat{m}_F}) + F(D_{\widehat{m}_F}) \leq P_n \gamma(\widehat{s}_m) + F(D_m) . \quad (7)$$

In particular, (7) holds for every $m \in \mathcal{M}_n$ such that $D_m = D_{\widehat{m}_F}$, for which $F(D_{\widehat{m}_F}) = F(D_m)$. Therefore, (7) implies that $\widehat{m}_F \in \mathcal{M}_{\text{dim}}(D_{\widehat{m}_F})$, hence $\widehat{m}_F \in \mathcal{M}_{\text{dim}}$. \square

Lemma 1 shows that despite the variety of functions F that can be used as a penalty, using a function of the dimensionality as a penalty always imply selecting among $(S_m)_{m \in \mathcal{M}_{\text{dim}}}$ (keeping in mind that \mathcal{M}_{dim} is random). Indeed, penalizing with a function of D means that all models of a given dimension D are penalized in the same way, so that the empirical risk alone is used for selecting among models of the same dimension. By extension, we will call *dimensionality-based model selection procedure* any procedure selecting a.s. $\widehat{m} \in \mathcal{M}_{\text{dim}}$.

Breiman [14] previously noticed that only a few models—called “RSS-extreme submodels”—can be selected by penalties of the form $F(D_m) = K D_m$ with $K \geq 0$. Although Breiman stated this limitation can be benefic from the computational point of view, results below show that this limitation precisely makes the quadratic risk increase when data are heteroscedastic.

4.3 Pros and cons of dimensionality-based model selection

As shown by equation (4), when data are heteroscedastic, $\mathbb{E}[\text{pen}_{\text{id}}(m)]$ is no longer proportional to the dimensionality D_m . The expectation of the ideal penalty actually is even not a function of D_m in general. Therefore, the unbiased risk estimation principle should prevent anyone from using dimensionality-based model selection procedures.

Nevertheless, dimensionality-based model selection procedures are still used for analyzing heteroscedastic data for at least three reasons:

- by ignorance of any other trustable model selection procedure than C_p , or of the assumptions of C_p ;
- because data are (wrongly) assumed to be homoscedastic;
- because they are simple and have a mild computational cost, no other measure of the size of the models being available.

The last two points can indeed be good reasons, provided that we know what we can loose—in terms of quadratic risk—by using a dimensionality-based model selection procedure instead of, for instance, some resampling-based penalty. The purpose of the next subsections is to estimate theoretically the price of violating the unbiased risk estimation principle in heteroscedastic regression.

4.4 Suboptimality of dimensionality-based penalization

Theorem 2 below shows that any dimensionality-based penalization procedure fails to attain asymptotic optimality for model selection among $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$ when data are heteroscedastic.

Theorem 2. *Assume that data $(X_1, Y_1), \dots, (X_n, Y_n) \in [0, 1] \times \mathbb{R}$ are independent and identically distributed (i.i.d.), X_i has a uniform distribution over \mathcal{X} and $\forall i = 1, \dots, n$, $Y_i = X_i + \sigma(X_i)\varepsilon_i$ where $(\varepsilon_i)_{1 \leq i \leq n}$ are independent such that $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathbb{E}[\varepsilon_i^2 | X_i] = 1$. Assume moreover that s is twice continuously differentiable,*

$$\|\varepsilon_i\|_\infty \leq E < \infty, \quad \min \left\{ (\sigma_a)^2, (\sigma_b)^2 \right\} > 0 \quad \text{and} \quad (\sigma_a)^2 \neq (\sigma_b)^2$$

$$\text{where} \quad (\sigma_a)^2 := \int_0^{1/2} (\sigma(x))^2 dx \quad \text{and} \quad (\sigma_b)^2 := \int_{1/2}^1 (\sigma(x))^2 dx.$$

Let $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, 1/2)}$ be the model collection defined in Section 2.3, with a maximal dimension $M_n = \lfloor n/(\ln(n))^2 \rfloor$. Then, constants $K_1, \mathbb{C}_1 > 0$ and an event of probability at least $1 - K_1 n^{-2}$ exist on which, for every function $F : \mathcal{M}_n \mapsto \mathbb{R}$ and every $\hat{m}_F \in \arg\min_{m \in \mathcal{M}_n} \{P_n \gamma(\hat{s}_m) + F(D_m)\}$,

$$\ell(s, \hat{s}_{\hat{m}_F}) \geq \mathbb{C}_1 \inf_{m \in \mathcal{M}_n} \{ \ell(s, \hat{s}_m) \} \quad \text{with} \quad \mathbb{C}_1 > 1. \quad (8)$$

The constant \mathbb{C}_1 may only depend on $(\sigma_a)^2 / (\sigma_b)^2$; the constant K_1 may only depend on $E, (\sigma_a)^2, (\sigma_b)^2, \|s'\|_\infty$ and $\|s''\|_\infty$.

Theorem 2 is proved in Section 7.4.

Remark 2. 1. The right-hand side of (8) is of order $n^{-2/3}$. Hence, no oracle inequality (2) for \hat{m}_F can be proved with a constant C tending to one when n tends to infinity and a remainder term $R_n \ll n^{-2/3}$.

2. Results similar to Theorem 2 can be proved similarly with other model collections (such as the nonregular ones defined Section 2.3) and with unbounded noises (thanks to concentration inequalities proved in [6]). The choice $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, 1/2)}$ in the statement of Theorem 2 only intends to keep the proof as simple as possible.

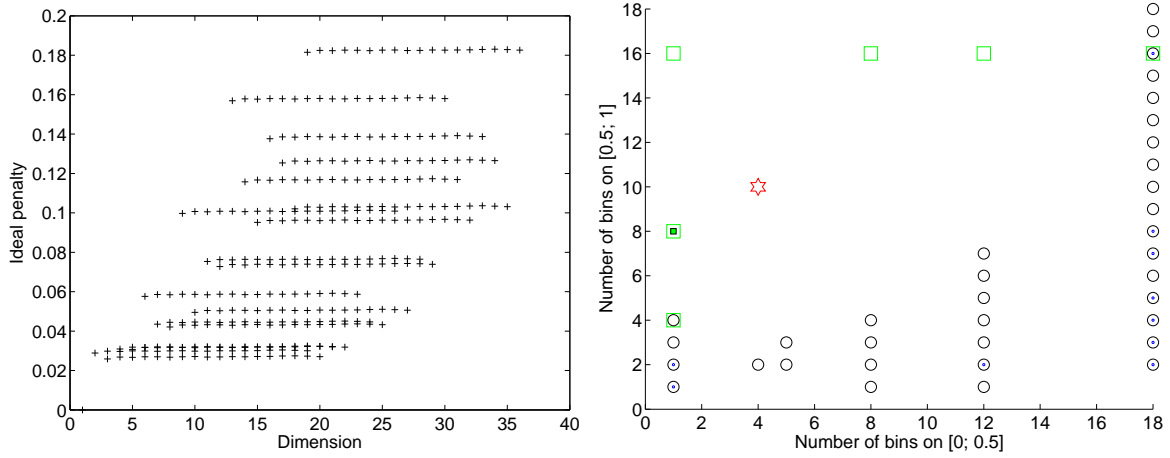


Figure 2: Experiment X1-005. Left: Ideal penalty *vs.* D_m for one particular sample. A similar picture holds in expectation. Right: The path of models that can be selected with penalties proportional to penLoo (green squares; the small square corresponds to pen=penLoo) is closest to the oracle (red star) than the path of models that can be selected with a dimensionality-based penalty (black circles; blue points correspond to linear penalties).

Theorem 2 is a quite strong result, implying that no dimensionality-based penalization procedure can satisfy an oracle inequality with leading constant $1 \leq C < \mathbb{C}_1$, even a procedure using the knowledge of s and σ ! The proof of Theorem 2 even shows that the *ideal dimensionality-based model selection procedure*, defined by

$$\hat{m}_{\text{dim}}^* \in \hat{m}(\hat{D}^*) \quad \text{where} \quad \hat{D}^* \in \operatorname{argmin}_{D \in \mathcal{D}_n} \{ \ell(s, \hat{s}_{\hat{m}(D)}) \} , \quad (9)$$

is suboptimal with a large probability.

The combination of Theorems 1 and 2 shows that when data are heteroscedastic, the price to pay for using a dimensionality-based model selection procedure instead of some resampling-based penalty is (at least) an increase of the quadratic risk by some multiplying factor $\mathbb{C}_1 > 1$ (except maybe for small sample sizes). Therefore, the computational cost of resampling has its counterpart in the quadratic risk. Empirical evidence for the same phenomenon in the context of multiple change-points detection can be found in [8].

4.5 Illustration of Theorem 2

Let us illustrate Theorem 2 and its proof with a simulation experiment, called ‘X1-005’: The model collection is $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$, and data $(X_i, Y_i)_{1 \leq i \leq n}$ are generated according to (1) with $s(x) = x$, $\sigma(x) = (1 + 19 \times \mathbf{1}_{x \leq 1/2})/20$, $X_i \sim \mathcal{U}([0, 1])$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $n = 200$ data points. An example of such data sample is plotted on the left of Figure 1, together with the oracle estimator \hat{s}_{m^*} .

Then, as remarked previously from (4), the ideal penalty is clearly not a function of the dimensionality (Figure 2 left). According to (4), the right penalty is not proportional to D_m but to $D_{m,1} \int_0^{1/2} \sigma^2(x) dx + D_{m,2} \int_{1/2}^1 \sigma^2(x) dx$. The consequence of this fact is that any $m \in \mathcal{M}_{\text{dim}}$ is far from the oracle, as shown by the right of Figure 2. Indeed, minimizing the empirical risk over models of a given dimension D leads to put more bins where the noise-level is larger, that is, to overfit locally (see also [8] for a deeper experimental study of this local overfitting

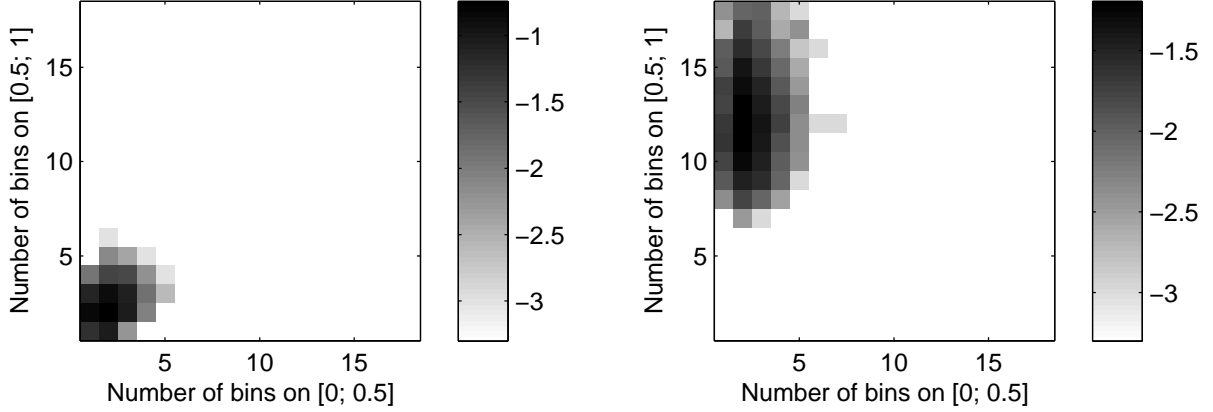


Figure 3: Experiment X1-005. Left: $\log_{10} \mathbb{P}(m = \hat{m}_{\text{lin}}^*)$ represented in \mathbb{R}^2 using $(D_{m,1}, D_{m,2})$ as coordinates, where \hat{m}_{dim}^* is defined by (9); $N = 10\,000$ samples have been simulated for estimating the probabilities. Right: $\log_{10} \mathbb{P}(m = m^*)$ using the same representation and the same $N = 10\,000$ samples.

phenomenon). Furthermore, Figure 3 shows that on 10 000 samples, m^* is almost always far from \mathcal{M}_{dim} , and in particular from \hat{m}_{dim}^* . On the contrary, using a resampling-based penalty (possibly multiplied by some factor $\mathcal{C}_{\text{ov}} > 0$) leads to avoid overfitting, and to select a model much closer to the oracle (see Figure 2 right).

4.6 Performance of linear penalties

Let us now focus on the most classical dimensionality-based penalties, that is, “linear penalties” of the form

$$\text{pen}(m) = KD_m \quad (10)$$

where $K > 0$ can be data-dependent, but does not depend on m . The first result of this subsection is that linear penalties satisfy an oracle inequality (2) (with leading constant $C > 1$) provided the constant K in (10) is large enough.

Proposition 2. *Assume that $(\text{AS})_{\text{hist}}$ holds true. Then, if*

$$\forall m \in \mathcal{M}_n, \quad \text{pen}(m) = \frac{KD_m}{n} \quad \text{with} \quad K > \|\sigma\|_{\infty}^2, \quad$$

constants $K_2, \mathbb{C}_2 > 0$ exist such that with probability at least $1 - K_2 n^{-2}$,

$$\ell(s, \hat{s}_{\hat{m}}) \leq \mathbb{C}_2 \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\} \quad (11)$$

The constant K_2 may depend on all the constants appearing in the assumptions (that is, $c_{\mathcal{M}}, \alpha_{\mathcal{M}}, c_{\text{rich}}, A, \sigma_{\min}, C_b^+, C_b^-, \beta_-, \beta_+, c_{r,\ell}^X$ and K ; assuming $K \geq 2\|\sigma\|_{\infty}^2$, K_2 does not depend on K). The constant \mathbb{C}_2 may only depend on K, σ_{\min} and $\|\sigma\|_{\infty}$; when $K \geq 2\|\sigma\|_{\infty}^2$, \mathbb{C}_2 can be made as close as desired to $K\sigma_{\min}^{-2} - 1$ at the price of enlarging K_2 .

Proposition 2 is proved in Section 7.5. As a consequence of Proposition 2, if we can afford loosing a constant factor of order $\|\sigma\|_{\infty}^2 / \sigma_{\min}^2$ in the quadratic risk, a relevant (and computationally cheap) strategy is the following: First, estimate an upper bound on $\|\sigma\|_{\infty}^2$. Second, plug

it into the penalty $\text{pen}(m) = K' \|\sigma\|_\infty^2 D_m/n$, where $K' > 1$ remains to be chosen. According to the proof of Proposition 2 and previous results in the homoscedastic case, $K' = 2$ should be a good choice in general. Let us now add a few comments.

Remark 3. 1. The assumption set $(\mathbf{AS})_{\text{hist}}$ is discussed in Remark 1 in Section 3.3, and can be relaxed in various ways, see [6].

2. Proposition 2 can be generalized to other models than sets of piecewise constant functions. Indeed, the keystone of the proof of Proposition 2 is that $2\sigma_{\min}^2 \leq nD_m^{-1} \mathbb{E}[\text{pen}_{\text{id}}(m)] \leq 2\|\sigma\|_\infty^2$, which holds for instance when the design is fixed and the models S_m are finite dimensional vector spaces. Therefore, using arguments similar to the ones of [12, 10] for instance, an oracle inequality like (11) can be proved when models are general vector spaces, assuming that the design $(X_i)_{1 \leq i \leq n}$ is deterministic.

The second result of this subsection shows that the condition $K > \|\sigma\|_\infty^2/n$ in Proposition 2 really prevents from strong overfitting. In particular, some example can be built where C_p strongly overfits.

Proposition 3. *Let us consider the framework of Section 2.1 with $\mathcal{X} = [0, 1]$, $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, 1/2)}$ with maximal dimension $M_n = \lfloor n/(\ln(n)) \rfloor$, and assume that:*

- (\mathbf{Ab}) and (\mathbf{An}) hold true (see the definition of $(\mathbf{AS})_{\text{hist}}$),
- $s \in \mathcal{H}(\alpha, R)$, that is, $\forall x_1, x_2 \in \mathcal{X}$, $|s(x_2) - s(x_1)| \leq R|x_2 - x_1|^\alpha$, for some $R > 0$ and $\alpha \in (0, 1]$,
- $\mu = \mathbb{P}(X \in [0, 1/2]) \in (0, 1)$,
- conditionally on $\{X \in [0, 1/2]\}$, X has a density w.r.t. the Lebesgue measure which is lower bounded by $c_{X, \text{Leb}} > 0$.

If in addition $\forall m \in \mathcal{M}_n$, $\text{pen}(m) = KD_m/n$ with $K < \inf_{t \in [0, 1/2]} \{\sigma(t)^2\}$, then constants $K_3, K_4, K_5 > 0$ exist such that with probability at least $1 - K_3 n^{-2}$,

$$D_{\hat{m}} \geq D_{\hat{m}, 1} \geq \frac{K_4 n}{\ln(n)} \quad (12)$$

$$\text{and } \ell(s, \hat{s}_{\hat{m}}) \geq \frac{K_5}{(\ln(n))^2} \geq \ln(n) \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\} . \quad (13)$$

The constants K_3, K_4, K_5 may depend on $A, \sigma_{\min}, \alpha, R, \mu, c_{X, \text{Leb}}, \inf_{[0, 1/2]} \{\sigma^2\}$ and K , but they do not depend on n .

Proposition 3, which is actually is a corollary of a more general result on minimal penalties—Theorem 2 in [9]—is proved in Section 7.6.

Consider in particular the following example: $X \in [0, 1]$ with a density w.r.t. Leb equal to $2\mu \mathbf{1}_{[0, 1/2]} + 2(1-\mu) \mathbf{1}_{(1/2, 1]}$ for some $\mu \in (0, 1)$ and $\sigma = \sigma_a \mathbf{1}_{[0, 1/2]} + \sigma_b \mathbf{1}_{(1/2, 1]}$ for some $\sigma_a \geq \sigma_b > 0$. Then, the penalty KD_m/n leads to overfitting as soon as $K < \sigma_a^2 = \|\sigma\|_\infty^2$, which shows that the lower bound $K > \|\sigma\|_\infty^2$ appearing in the proof of Proposition 2 cannot be improved in this example.

Let us now consider C_p , that we naturally generalize to the heteroscedastic case by $\text{pen}_{\text{Cp}}(m) = KD_m/n$ with $K = 2\mathbb{E}[\sigma(X)^2]$. In the above example, $K = 2\mu\sigma_a^2 + 2(1-\mu)\sigma_b^2$. So, the condition on K in Proposition 3 can be written

$$\mu + (1-\mu) \frac{\sigma_b^2}{\sigma_a^2} < \frac{1}{2} ,$$

Table 1: Parameters of the four experiments.

Experiment	X1-005	S0-1	XS1-05	X1-005 μ 02
$s(x)$	x	$\sin(\pi x)$	see Eq. (15)	x
$\sigma(x)$	$(1 + 19 \times \mathbf{1}_{x \leq 1/2})/20$	$\mathbf{1}_{x > 1/2}$	$(1 + \mathbf{1}_{x \leq 1/2})/2$	$(1 + 19 \times \mathbf{1}_{x \leq 1/2})/20$
n	200	200	500	1000
$\mathbb{P}(X \leq 1/2)$	1/2	1/2	1/2	1/5
M_n	$\lfloor n/(\ln(n)) \rfloor = 37$	$\lfloor n/(\ln(n)) \rfloor = 37$	$\lfloor n/(\ln(n)) \rfloor = 80$	$\lfloor n/(\ln(n))^2 \rfloor = 20$

which holds when

$$\mu \in (0, 1/2) \quad \text{and} \quad \frac{\sigma_b^2}{\sigma_a^2} < \frac{\frac{1}{2} - \mu}{1 - \mu} . \quad (14)$$

Therefore, when (14) holds, Proposition 3 shows that C_p strongly overfits.

The conclusion of this subsection is that some linear penalties can be used with heteroscedastic data provided that we can estimate $\|\sigma\|_\infty^2$ by some $\widehat{\sigma}_\infty^2$ such that $2\widehat{\sigma}_\infty^2 > \|\sigma\|_\infty^2$ holds with probability close to 1. Then the price to pay is an increase of the quadratic risk by a constant factor of order $\max(\sigma^2)/\min(\sigma^2)$ in general.

5 Simulation study

This section intends to compare by a simulation study the finite sample performances of the model selection procedures studied in the previous sections: dimensionality-based and resampling-based procedures.

5.1 Experiments

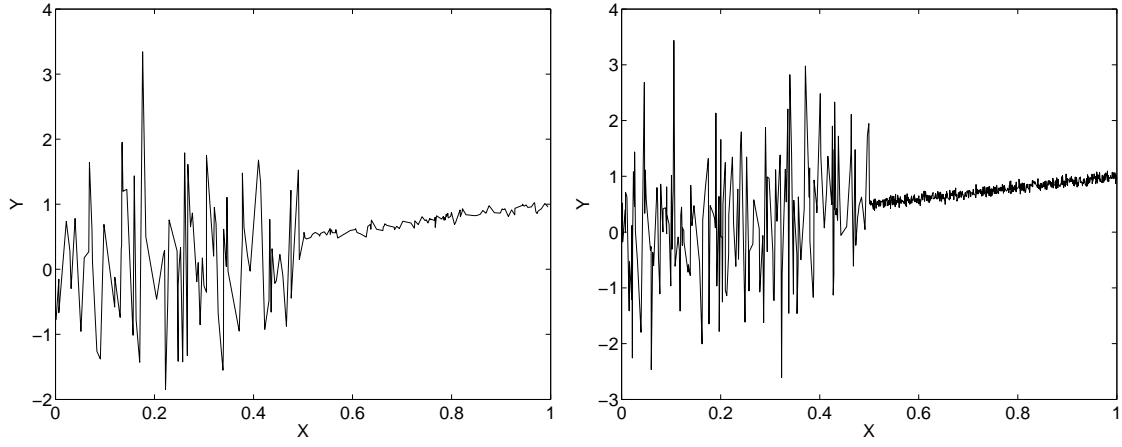
We consider four experiments, called ‘X1-005’ (as in Section 4.5), ‘X1-005 μ 02’, ‘S0-1’ and ‘XS1-05’. Data $(X_i, Y_i)_{1 \leq i \leq n}$ are generated according to (1) with $\varepsilon_i \sim \mathcal{N}(0, 1)$ and X_i has density w.r.t. $\text{Leb}([0, 1])$ of the form $2\mu\mathbf{1}_{[0, 1/2]} + 2(1 - \mu)\mathbf{1}_{(1/2, 1]}$, where $\mu = \mathbb{P}(X_1 \leq 1/2) \in (0, 1)$. The functions s and σ , and the values of n and μ , depend on the experiment, see Table 1 and Figure 4; in experiment ‘XS1-05’, the regression function is given by

$$s(x) = \frac{x}{4}\mathbf{1}_{x \leq 1/2} + \left[\frac{1}{8} + \frac{2}{3} \sin(16\pi x) \right] \mathbf{1}_{x > 1/2} . \quad (15)$$

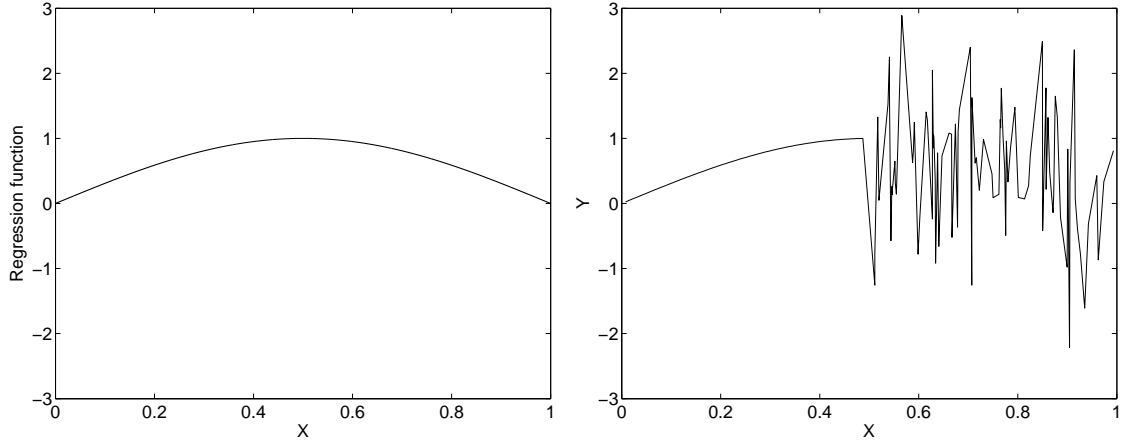
In each experiment, $N = 10\,000$ independent data samples are generated, and the model collection is $(S_m)_{m \in \mathcal{M}_n^{(\text{reg}, 1/2)}}$, with different values of M_n for computational reasons, see Table 1. The signal-to-noise ratio is rather small in the four experimental settings considered here, and the collection of models is quite large ($\text{Card}(\mathcal{M}_n) = 1 + (M_n/2)^2$). Therefore, we can expect overpenalization to be necessary (see Section 6.3.2 of [6] for more details on overpenalization).

5.2 Procedures compared

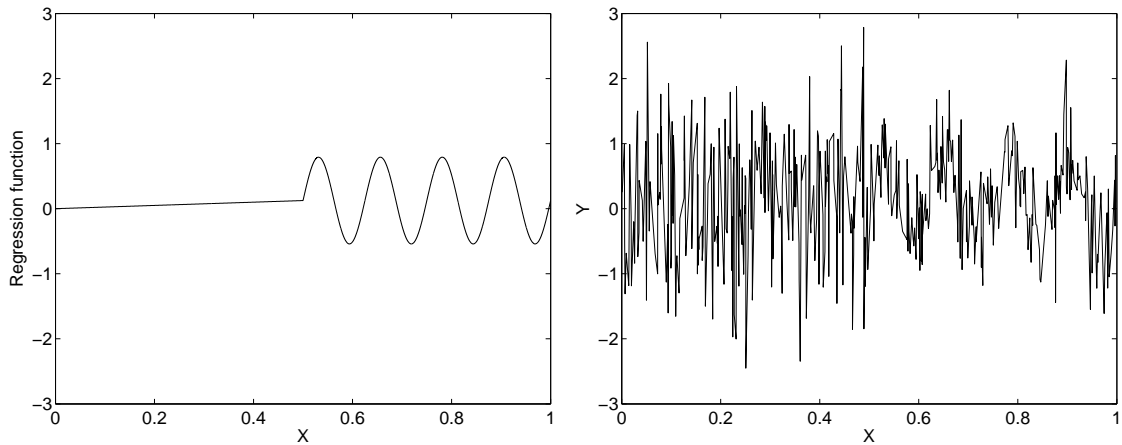
For each sample, the following model selection procedures are compared, where τ denotes a permutation of $\{1, \dots, n\}$ such that $(X_{\tau(i)})_{1 \leq i \leq n}$ is nondecreasing.



Experiments X1-005 (left) and X1-005 μ 02 (right): one data sample



Experiment S0-1 (left: regression function; right: one data sample)



Experiment XS1-05 (left: regression function; right: one data sample)

Figure 4: Regression functions and one particular data sample for the four experiments.

(A) Epenid: penalization with

$$\text{pen}(m) = \mathbb{E} [\text{pen}_{\text{id}}(m)] \quad \text{where} \quad \text{pen}_{\text{id}}(m) = P\gamma(\widehat{s}_m) - P_n\gamma(\widehat{s}_m) ,$$

as defined in Section 2.2. This procedure makes use of the knowledge of the true distribution P of data. Its model selection performances witness what performances could be expected (ideally) from penalization procedures adapting to heteroscedasticity.

(B) MalEst: penalization with $\text{pen}_{\text{Cp}}(m)$ where the variance is estimated as in Section 6 of [10], that is,

$$\text{pen}(m) = \frac{2\widehat{\sigma}^2 D_m}{n} \quad \text{with} \quad \widehat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n/2} (Y_{\tau(2i)} - Y_{\tau(2i-1)})^2 ,$$

Replacing $\widehat{\sigma}^2$ by $\mathbb{E} [\sigma(X)^2]$ doesn't change much the performances of MalEst, see Appendix A.

(C) MalMax: penalization with $\text{pen}(m) = 2 \|\sigma\|_\infty^2 D_m/n$ (using the knowledge of σ).

(D) HO: hold-out procedure, that is,

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}_n} \left\{ P_n^{(I^c)} \gamma \left(\widehat{s}_m^{(I)} \right) \right\} ,$$

where $P_n^{(I^c)}$ and $\widehat{s}^{(I)}$ are defined as in Section 3.3, and $I \subset \{1, \dots, n\}$ is uniformly chosen among subsets of size $n/2$ such that $\forall k \in \{1, \dots, n/2\}$, $\text{Card}(I \cap \{\tau(2k-1), \tau(2k)\}) = 1$.

(E–F–G) VFCV (V -fold cross-validation) with $V = 2, 5$ and 10 :

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}_n} \left\{ \frac{1}{V} \sum_{j=1}^V P_n^{(B_j)} \gamma \left(\widehat{s}_m^{(B_j^c)} \right) \right\} ,$$

where $(B_j)_{1 \leq j \leq V}$ is a regular partition of $\{1, \dots, V\}$, uniformly chosen among partitions such that $\forall j \in \{1, \dots, V\}$, $\forall k \in \{1, \dots, n/V\}$, $\text{Card}(B_j \cap \{\tau(i) \text{ s.t. } kV - V + 1 \leq i \leq kV\}) = 1$.

(H) penHO: hold-out penalization, as defined in Section 3.3, with the same training set I as in procedure (D).

(I–J–K) penVF (V -fold penalization) with $V = 2, 5$ and 10 , as defined by (5), with the same partition $(B_j)_{1 \leq j \leq V}$ as in procedures (E–F–G) respectively.

(L) penLoo (Leave-one-out penalization), that is, V -fold penalization with $V = n$ and $\forall j$, $B_j = \{j\}$.

Every penalization procedure was also performed with various overpenalization factors $\mathcal{C}_{\text{ov}} \geq 1$, that is, with $\text{pen}(m)$ replaced by $\mathcal{C}_{\text{ov}} \times \text{pen}(m)$. Only results with $\mathcal{C}_{\text{ov}} \in \{1, 2, 4\}$ are reported in the paper since they summarize well the whole picture.

Table 2: Short names for the procedures compared.

A: Epenid	E: 2-FCV	I: pen2-F
B: MalEst	F: 5-FCV	J: pen5-F
C: MalMax	G: 10-FCV	K: pen10-F
D: HO	H: penHO	L: penLoo

Furthermore, given each penalization procedure among the above (let us call it ‘Pen’), we consider the associated ideally calibrated penalization procedure ‘IdPen’, which is defined as follows:

$$\begin{aligned} \hat{m}_{\text{pen}}^* = \hat{m}_{\text{pen}} \left(\hat{K}_{\text{pen}}^* \right) \quad \text{where} \quad \forall K \geq 0, \quad \hat{m}_{\text{pen}}(K) \in \arg \min_{m \in \mathcal{M}_n} \{ P_n \gamma(\hat{s}_m) + K \text{pen}(m) \} \\ \text{and} \quad \hat{K}_{\text{pen}}^* \in \arg \min_{K \geq 0} \{ \ell(s, \hat{s}_{\hat{m}_{\text{pen}}(K)}) \} . \end{aligned}$$

In other words, $\text{pen}(m)$ is used with the best distribution and data-dependent overpenalization factor \hat{K}_{pen}^* . Needless to say, ‘IdPen’ makes use of the knowledge of P , and is only considered for experimental comparison.

When $\text{pen}(m) = D_m$, the above definition defines the *ideal linear penalization procedure*, that we call ‘IdLin’ (and the selected model is denoted by \hat{m}_{lin}^*). In addition, we consider the ideal dimensionality-based model selection procedure ‘IdDim’, defined by (9).

Finally, let us precise that in all the experiments, prior to performing any model selection procedure, models S_m such that $\min_{\lambda \in \Lambda_m} \text{Card} \{ i \text{ s.t. } X_i \in \lambda \} < 2$ are removed from $(S_m)_{m \in \mathcal{M}_n}$. Without removing interesting models, this preliminary step intends to provide a fair and clear comparison between penLoo (which was defined in [6] including this preliminary step) and other procedures.

The benchmark for comparing model selection performances of the procedures is

$$C_{\text{or}} := \frac{\mathbb{E}[\ell(s, \hat{s}_{\hat{m}})]}{\mathbb{E}[\inf_{m \in \mathcal{M}_n} \ell(s, \hat{s}_m)]} , \quad (16)$$

where both expectations are approximated by an average over the N simulated samples. Basically, C_{or} is the constant that should appear in an oracle inequality (2) holding in expectation with $R_n = 0$. We also report the following uncertainty measure of our estimator of C_{or} ,

$$\varepsilon_{C_{\text{or}}, N} := \frac{\sqrt{\text{var}(\ell(s, \hat{s}_{\hat{m}}))}}{\sqrt{N} \mathbb{E}[\inf_{m \in \mathcal{M}_n} \ell(s, \hat{s}_m)]} , \quad (17)$$

where var (resp. \mathbb{E}) is approximated by an empirical variance (resp. expectation) over the N simulated samples.

5.3 Results

The (evaluated) values of $C_{\text{or}} \pm \varepsilon_{C_{\text{or}}, N}$ in the four experiments are given on Figures 5 and 6 (for procedures A–L) and in Table 3 (for IdDim and the ideally calibrated penalization procedures). In addition, results for experiment ‘X1–005’ with various values of the sample size n are presented on Figure 7. Note that a few additional results are provided in Appendix A.

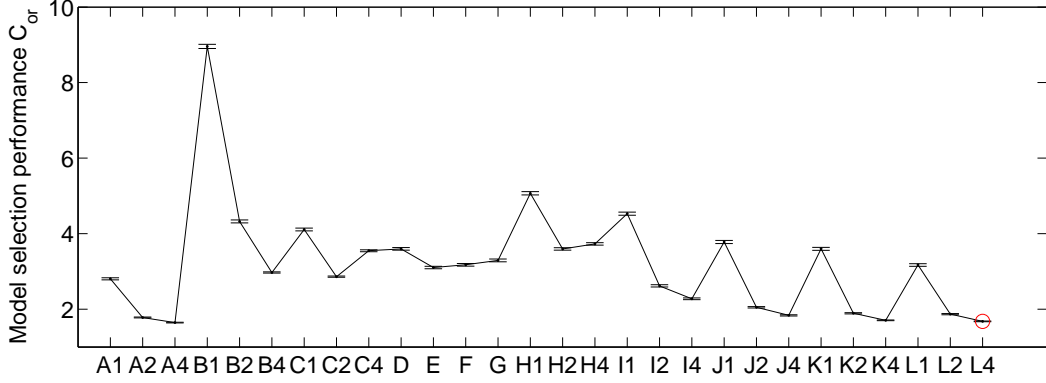


Figure 5: Accuracy indices C_{or} in Experiment X1–005 for various model selection procedures. Error bars represent $\varepsilon_{C_{or},N}$. C_{or} is defined by (16) and $\varepsilon_{C_{or},N}$ by (17). Red circles show which are the most accurate procedures, taking uncertainty into account and excluding procedures using $\mathbb{E}[\text{pen}_{id}]$. On the x-axis, procedures are named with a letter (whose meaning is given in Table 2), plus a figure (for penalization procedures) equal to the value of the overpenalization constant C_{ov} ; for instance, J2 means the 5-fold penalty multiplied by 2. See also Table 4 and Figures 9–10 in Appendix A.

Table 3: Accuracy indices C_{or} for “ideal” procedures in four experiments, $\pm \varepsilon_{C_{or},N}$. See also Figures 11–12 in Appendix A.

Experiment	X1–005	S0–1	XS1–05	X1–005 μ 02
IdLin	2.065 ± 0.010	2.106 ± 0.009	1.308 ± 0.002	2.211 ± 0.009
IdDim	1.507 ± 0.009	1.595 ± 0.008	1.262 ± 0.002	1.683 ± 0.008
IdPenHO	2.158 ± 0.020	1.785 ± 0.012	1.509 ± 0.005	1.767 ± 0.012
IdPen2F	1.454 ± 0.011	1.523 ± 0.009	1.303 ± 0.003	1.410 ± 0.008
IdPen5F	1.377 ± 0.008	1.467 ± 0.008	1.244 ± 0.002	1.413 ± 0.007
IdPen10F	1.384 ± 0.008	1.446 ± 0.008	1.240 ± 0.002	1.414 ± 0.007
IdPenLoo	1.378 ± 0.008	1.458 ± 0.008	1.233 ± 0.002	1.419 ± 0.007
IdEpenid	1.363 ± 0.008	1.401 ± 0.008	1.232 ± 0.002	1.410 ± 0.007

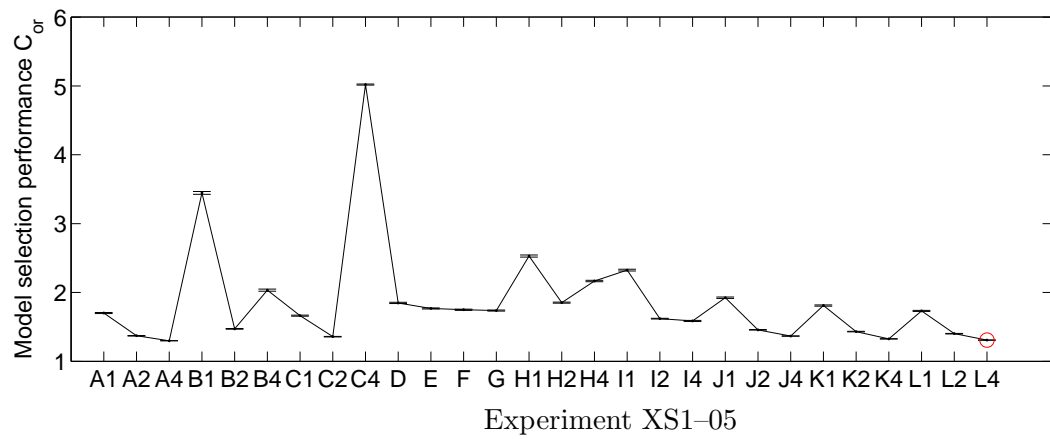
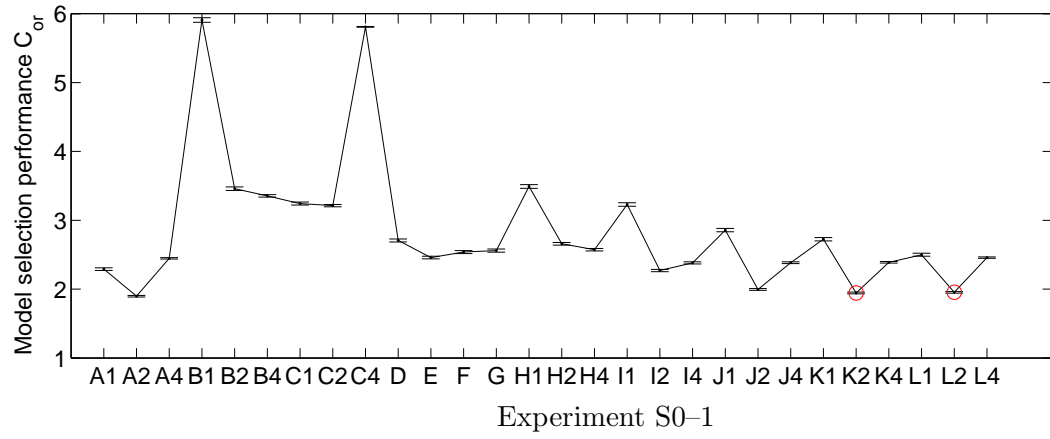
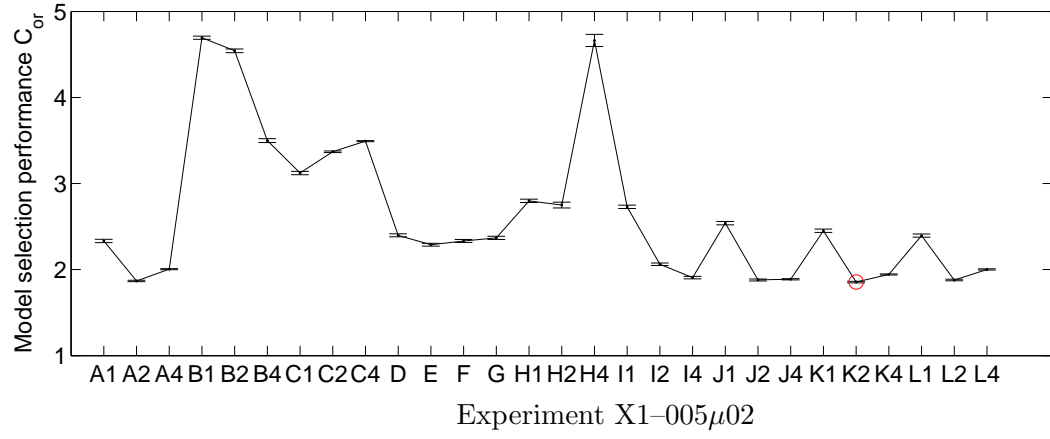


Figure 6: Same as Figure 5 with the three other experiments. See also Table 4 and Figures 9–10 in Appendix A.

The first remark we can make from Figures 5 and 6 is that changing the overpenalization factor \mathcal{C}_{ov} can make large differences in C_{or} , as it was pointed in [5]. We do not address here the question of choosing \mathcal{C}_{ov} from data since it is beyond the scope of the paper; see Section 11.3.3 of [4] for suggestions.

The choice of the overpenalization factor can be put aside in two ways. First, we can compare on Figures 5 and 6 the performances obtained with the best deterministic value of \mathcal{C}_{ov} for each procedure. Second, we can compare procedures with their best data-driven (and distribution dependent) overpenalization factor, as done in Table 3. Both ways yield the same qualitative conclusion in the four experiments—up to minor variations—which can be summarized as follows.

Firstly, most resampling-based procedures (that is, VFCV, penVF, penLoo) outperform dimensionality-based procedures (MalEst, which strongly overfits, and even MalMax), which confirms the theoretical results of Sections 3 and 4. In particular, penLoo yields a significant improvement over MalEst and MalMax (by more than 25% in three experiments, and by 2.3% in XS1–05), and IdPenLoo similarly outperforms IdLin and IdDim (by more than 8.5% in three experiments, and by 2.3% in XS1–05). Even penLoo with a well-chosen *deterministic* overpenalization factor \mathcal{C}_{ov} outperforms IdLin by 7% to 18% in experiments X1–005, S0–1 and X1–005 μ 02, and penLoo equals the performances of IdLin in experiment XS1–05 (compare Table 3 with Table 4 in Appendix A). Figure 7 illustrates the same phenomenon: when the sample size increases, the model selection performance of IdLin remains approximately constant (close to 2) while the model selection performance C_{or} of penLoo constantly decreases (with $\mathcal{C}_{\text{ov}} = 1.25$ because overpenalization is still needed for $n = 3000$ and we could not consider larger sample sizes for computational reasons). The reasons for this clear advantage of resampling-based procedures are the same in the four experiments: As pointed out in Section 4.5 for experiment X1–005, no dimensionality-based model selection procedure can select a model close enough to the oracle m^* . In particular, figures similar to Figure 3 hold in experiments S0–1, X1–005 μ 02, and XS1–05, see Figure 13 in Appendix A.

Secondly, improving over dimensionality-based procedures requires a significant increase of the computational cost. Indeed, PenHO performs significantly worse than MalMax in experiments X1–005 and XS1–05, while penHO and MalMax have similar performances in experiments S0–1 and X1–005 μ 02. Furthermore, IdPenHO performs worse than IdDim in the four experiments, and even worse than IdLin in experiments X1–005 and XS1–05. In order to obtain sensibly better performances than dimensionality-based procedures, our experiments show that the computational cost must at least be increased to the one of 5-fold penalization. This phenomenon certainly comes from the small signal-to-noise ratio, which makes it difficult to estimate precisely the penalty shape by resampling, whereas MalMax can provide reasonably good performances thanks to underfitting.

Finally, let us add that penVF outperforms VFCV (and similarly penHO outperforms HO), provided the (deterministic) overpenalization factor is well-chosen, as shown in a previous paper [5].

6 Conclusion: How to choose the penalty?

Combining the theoretical results of Sections 3 and 4 with the conclusions of the experiments of Section 5, we can propose an answer to the main question raised in this paper: Which penalty should be used for which model selection problem? A visual summary of this conclusion is proposed on Figure 8.

Three main factors must be taken into account to answer this question:

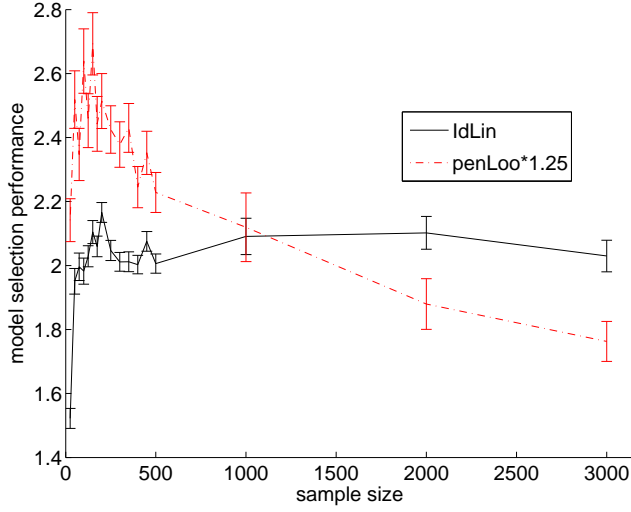


Figure 7: Model selection performance C_{or} of IdLin and penLoo (multiplied by $C_{ov} = 1.25$) as a function of n for experiment X1-005. Each estimated value of C_{or} is obtained with $N = 1000$ data samples for $n \leq 500$ and $N = 200$ data samples for $n \geq 1000$ for computational reasons. Error bars represent $\varepsilon_{C_{or}, N}$.

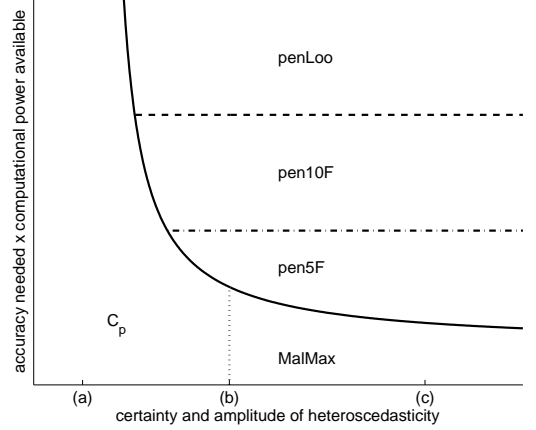


Figure 8: Summary of Section 6: best penalty as a function of the computational power available, the level of accuracy needed, and the knowledge about heteroscedasticity of data. The limit between linear and resampling penalties goes up when the SNR goes down.

1. the prior knowledge on the noise-level $\sigma(\cdot)$,
2. the trade-off between computational power and statistical accuracy desired,
3. the signal-to-noise ratio (SNR).

What is known about $\sigma(\cdot)$ appears as the determinant factor:

- (a) If $\sigma(\cdot)$ is known to be constant, C_p clearly is the best procedure, compared to cross-validation or resampling-based procedures which cannot take into account this information about data.
- (d) If $\sigma(\cdot)$ is non-constant but completely known, then the expectation of the ideal penalty $\mathbb{E}[\text{pen}_{id}(m)]$ is entirely known and should be used, following the unbiased risk estimation principle.

Note that C_p or AIC are still often used in case (d), mainly by non-statisticians that probably do not know (or do not trust) model selection procedures adapting to heteroscedasticity. This paper provides clear theoretical arguments to show them what improvement they could obtain by using a properly chosen procedure.

Choosing a penalty is less simple in the following two intermediate cases, where a trade-off must be found between the precise knowledge on σ , computational power and statistical accuracy:

- (b) $\sigma(\cdot)$ is probably (almost) constant, but this information is questionable
- (c) $\sigma(\cdot)$ is known to be non-constant, without prior information on its shape

If the computational power has no limits (or, equivalently, if accuracy is crucial), resampling penalties (or V -fold penalties with V large) should be used in both cases.

Nevertheless, when the computational power is limited, one has to take into account that V -fold penalties with too small V poorly estimate the shape of the penalty, so that they may be outperformed by MalMax (that is, C_p with σ^2 replaced by some upper-bound on $\|\sigma\|_\infty^2$), in particular in case (b). Similarly, if the final user does not matter losing a (small) constant in the quadratic risk, MalMax could be used instead of resampling in case (b), and even in case (c) provided $(\max \sigma / \min \sigma)^2$ is small enough. Of course, using MalMax requires either the knowledge of an upper bound on $\|\sigma\|_\infty^2$, or to be able to estimate one (for instance assuming $\sigma(\cdot)$ does not jump too much).

The picture can also change depending on the SNR. When the SNR is small, overpenalization is usually required. Therefore, choosing a proper overpenalization level can be more important than estimating the shape of the penalty, so that MalMax (possibly with an enlarged penalty) is quite a reasonable choice in case (b), and even in case (c) depending on the computational power. On the contrary, when the SNR is large, V -fold penalties (even with rather small V , such as $V = 5$) yield a significant improvement over any dimensionality-based penalty.

A natural question arises from this conclusion: How to calibrate precisely the constant in front of the penalty? Birgé and Massart [13] proposed an optimal (and computationally cheap) data-driven procedure answering this question, based upon the concept of minimal penalties (see [9] for the heteroscedastic regression framework). Nevertheless, theoretical results on Birgé and Massart's procedure are not accurate enough to determine whether it takes into account the need for overpenalization when the SNR is small. Therefore, understanding precisely how we should overpenalize as a function of the SNR seems a quite important question from the practical point of view, which is still widely open, up to the best of our knowledge.

7 Proofs

Before proving Theorem 2 and Propositions 2 and 3, let us define some notation and recall probabilistic results from other papers [6, 5, 9] that are used in the proof.

7.1 Notation

In the rest of the paper, L denotes an absolute constant, not necessarily the same at each occurrence. When L is not universal, but depends on p_1, \dots, p_k , it is written L_{p_1, \dots, p_k} .

Define, for every model $m \in \mathcal{M}_n$,

$$\begin{aligned} p_1(m) &:= P\gamma(\hat{s}_m) - P\gamma(s_m) & p_2(m) &:= P_n\gamma(s_m) - P_n\gamma(\hat{s}_m) \quad \text{and} \\ \bar{\delta}(m) &:= (P_n - P)(\gamma(s_m) - \gamma(s)) \quad . \end{aligned}$$

7.2 Probabilistic tools: expectations

Proposition 4 (Proposition 1 and Lemma 7 in [5]). *Let S_m be the model of histograms associated with the partition Λ_m . Then,*

$$\mathbb{E}[p_1(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} (1 + \delta_{n,p_\lambda}) \sigma_\lambda^2 \quad (18)$$

$$\mathbb{E}[p_2(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \sigma_\lambda^2 \quad (19)$$

$$\text{where } \sigma_\lambda^2 := \mathbb{E} \left[(Y - s(X))^2 \mid X \in \lambda \right] = \left(\sigma_\lambda^d \right)^2 + \left(\sigma_\lambda^r \right)^2 ,$$

$p_\lambda = \mathbb{P}(X \in \lambda)$ and $\delta_{n,p}$ only depends on (n, p) . Moreover, $\delta_{n,p}$ is small when the product np is large:

$$|\delta_{n,p}| \leq \min \left\{ L_1, L_2(np)^{-1/4} \right\} ,$$

where L_1 and L_2 are absolute constants.

Note that $\delta_{n,p}$ can be made explicit: $\delta_{n,p} = np\mathbb{E}[Z^{-1}\mathbf{1}_{Z>0}] - 1$ where Z is a binomial random variable with parameters (n, p) .

Remark 4. The regressogram estimator \hat{s}_m is not defined when $\text{Card}\{i \text{ s.t. } X_i \in \lambda\} = 0$ for some $\lambda \in \Lambda_m$, which occurs with positive probability. Therefore, a convention for $p_1(m)$ as to be chosen on this event (which has a small probability, see Claim 1 in the proof of Theorem 2) so that $p_1(m)$ has a finite expectation (see [5] for details). This convention is purely formal, since the statement of Theorem 2 does not involve the expectation of $p_1(m)$. The important point is that the same convention is used in Proposition 5 below.

7.3 Probabilistic tools: concentration inequalities

We state in this section some concentration results on the components of the ideal penalty, using for $p_1(m)$ the same convention as in Proposition 4.

Proposition 5 (Proposition 10 in [6], proved in Section 4 of [7]). *Let $\gamma > 0$. Assume that $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n \geq 1$, $\|Y\|_\infty \leq A < \infty$ and*

$$D_m^{-1} \sum_{\lambda \in \Lambda_m} \mathbb{E}[\sigma(X)^2 \mid X \in \lambda] \geq Q > 0 .$$

Then, an event of probability at least $1 - Ln^{-\gamma}$ exists on which

$$p_1(m) \geq \mathbb{E}[p_1(m)] - L_{A,Q,\gamma} \left[(\ln n)^2 D_m^{-1/2} + e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad (20)$$

$$p_1(m) \leq \mathbb{E}[p_1(m)] + L_{A,Q,\gamma} \left[(\ln n)^2 D_m^{-1/2} + \sqrt{D_m} e^{-LB_n} \right] \mathbb{E}[p_2(m)] \quad (21)$$

$$|p_2(m) - \mathbb{E}[p_2(m)]| \leq L_{A,Q,\gamma} D_m^{-1/2} \ln(n) \mathbb{E}[p_2(m)] . \quad (22)$$

Lemma 6 (Proposition 8 in [9]). *Assume that $\|Y\|_\infty \leq A < \infty$. Then for any $x \geq 0$, an event of probability at least $1 - 2e^{-x}$ exists on which*

$$\forall \eta > 0, \quad |\bar{\delta}(m)| \leq \eta \ell(s, s_m) + \left(\frac{4}{\eta} + \frac{8}{3} \right) \frac{A^2 x}{n} . \quad (23)$$

Lemma 7 (Lemma 12 in [5]). *Let $(p_\lambda)_{\lambda \in \Lambda_m}$ be non-negative real numbers of sum 1, $(n\hat{p}_\lambda)_{\lambda \in \Lambda_m}$ a multinomial vector of parameters $(n; (p_\lambda)_{\lambda \in \Lambda_m})$, and $\gamma > 0$. Assume that $\text{Card}(\Lambda_m) \leq n$ and $\min_{\lambda \in \Lambda_m} \{np_\lambda\} \geq B_n > 0$. Then, an event of probability at least $1 - Ln^{-\gamma}$ exists on which*

$$\min_{\lambda \in \Lambda_m} \{n\hat{p}_\lambda\} \geq \frac{\min_{\lambda \in \Lambda_m} \{np_\lambda\}}{2} - 2(\gamma + 1) \ln n . \quad (24)$$

7.4 Proof of Theorem 2

In the following, $L_{(\mathbf{HThm})} = L_{E, \sigma_a, \sigma_b, \|\sigma'\|_\infty, \|\sigma''\|_\infty}$ denotes any constant depending on $E, \sigma_a, \sigma_b, \|\sigma'\|_\infty$ and $\|\sigma''\|_\infty$ only. The outline of the proof of Theorem 2 is the following:

- From Lemma 1, it is sufficient to prove that for every D ,

$$\inf_{m \in \mathcal{M}_{\dim(D)}} \{\ell(s, \hat{s}_m)\} \geq \mathbb{C}_1 \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\}$$

holds with probability at least $1 - K_1 n^{-2}$, with $\mathbb{C}_1 > 1$.

- Prove that all regressogram estimators are well-defined with a large probability (Claim 1).
- Compute explicitly the bias of each models (Claim 2).
- Provide a good approximation of the excess loss and the empirical risk of each model on a large probability event (Claim 3).
- Upper bound the excess loss of the oracle model (Claim 4).
- Lower bound the excess loss of small models (Claim 5).
- Prove that all models m having an excess loss close to the one of the oracle are close to the oracle model (Claim 6).
- Conclude the proof by showing that for every model close to the oracle, a model with a smaller empirical risk can be found.

As pointed out by Remark 4, the regressogram estimator associated with model S_m is not well defined if for some $\lambda \in \Lambda_m$, no X_i belongs to λ . The following claim shows this only happens with a small probability, hence this possible problem can be put aside for proving Theorem 2.

Claim 1. *An event Ω_1 of probability at least $1 - Ln^{-2}$ exists on which*

$$\forall m \in \mathcal{M}_n, \forall \lambda \in \Lambda_m, \quad \text{Card} \{i \text{ s.t. } X_i \in \lambda\} \geq 1 .$$

Hence, on Ω_1 , all estimators $(\hat{s}_m)_{m \in \mathcal{M}_n}$ are well-defined.

proof of Claim 1. For every $m \in \mathcal{M}_n$, let us apply Lemma 7 with $B_n = (\ln(n))^2$ and $\gamma = 4$. An event of probability at least $1 - Ln^{-4}$ exists such that

$$\min_{\lambda \in \Lambda_m} \text{Card} \{i \text{ s.t. } X_i \in \lambda\} \geq \frac{(\ln(n))^2}{2} - 10 \ln n .$$

This lower bound is positive provided that $n \geq L$. Therefore, the result holds on the intersection Ω_1 of these $\text{Card}(\mathcal{M}_n) \leq n^2$ events. \square

The next step is to use the results recalled in Sections 7.2 and 7.3 in order to control the excess loss and the empirical risk of each model. This leads to Claims 2 and 3 below.

Claim 2. *Define*

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} (s'(x))^2 dx \quad \text{and} \quad \alpha_2 = \frac{1}{48} \int_{1/2}^1 (s'(x))^2 dx .$$

For every $m \in \mathcal{M}_n$, some $\kappa_{m,b,1}, \kappa_{m,b,2} \in \mathbb{R}$ exist such that

$$\ell(s, s_m) = \alpha_1 D_{m,1}^{-2} (1 + \kappa_{m,b,1}) + \alpha_2 D_{m,2}^{-2} (1 + \kappa_{m,b,2}) \quad (25)$$

and $|\kappa_{m,b,i}| \leq L_{\|s'\|_\infty, \|s''\|_\infty} D_{m,i}^{-1}$.

proof of Claim 2. Since X is uniformly distributed on $[0, 1]$,

$$\ell(s, s_m) = \int_0^1 (s_m(X) - s(X))^2 = \sum_{\lambda \in \Lambda_m} \int_\lambda (s(x) - s_\lambda)^2 dx \quad (26)$$

where s_λ is the average of s on λ . We now fix some $\lambda \in \Lambda_m$. Let c_λ denote the center of the interval λ , and $|\lambda|$ the length of λ . Then,

$$\int_\lambda (s(x) - s_\lambda)^2 dx = (s(c_\lambda) - s_\lambda)^2 + \int_\lambda (s(x) - s(c_\lambda))^2 dx . \quad (27)$$

In addition, since s is twice continuously differentiable, for every $x \in \lambda$, some $g(x) \in \lambda$ exists such that

$$s(x) - s(c_\lambda) = (x - c_\lambda)s'(c_\lambda) + \frac{1}{2}(x - c_\lambda)^2 s''(g(x)) . \quad (28)$$

On the one hand, integrating (28) over λ leads to

$$(s_\lambda - s(c_\lambda))^2 \leq L \|s''\|_\infty^2 |\lambda|^4 . \quad (29)$$

On the other hand, integrating the square of (28) over λ leads to

$$\left| \int_\lambda (s(x) - s(c_\lambda))^2 dx - \frac{s'^2(c_\lambda) |\lambda|^3}{12} \right| \leq L |\lambda|^4 \|s''\|_\infty (\|s'\|_\infty + \|s''\|_\infty) . \quad (30)$$

Combining (27) with (29) and (30) then shows that for every $\lambda \in \Lambda_m$,

$$\left| \int_\lambda (s(x) - s_\lambda)^2 dx - \frac{s'^2(c_\lambda) |\lambda|^3}{12} \right| \leq L |\lambda|^4 \|s''\|_\infty (\|s'\|_\infty + \|s''\|_\infty) . \quad (31)$$

Furthermore, for every $\lambda \in \Lambda_m$,

$$\begin{aligned} \left| |\lambda| (s'(c_\lambda))^2 - \int_\lambda (s'(x))^2 dx \right| &\leq \int_\lambda \left| (s'(x))^2 - (s'(c_\lambda))^2 \right| dx \leq 2 \|s'\|_\infty \int_\lambda |s'(x) - s'(c_\lambda)| dx \\ &\leq 2 \|s'\|_\infty \|s''\|_\infty |\lambda|^2 . \end{aligned} \quad (32)$$

Using (26), combining (32) with (31) and summing over $\lambda \in \Lambda_m$ implies

$$\begin{aligned} \left| \ell(s, s_m) - \frac{1}{12} \sum_{\lambda \in \Lambda_m} \left(|\lambda|^2 \int_\lambda (s'(x))^2 dx \right) \right| &\leq L \|s''\|_\infty (\|s'\|_\infty + \|s''\|_\infty) \sum_{\lambda \in \Lambda_m} |\lambda|^4 \\ &\leq L \|s''\|_\infty (\|s'\|_\infty + \|s''\|_\infty) (D_{m,1}^{-3} + D_{m,2}^{-3}) \end{aligned}$$

and the result follows. \square

Claim 3. Define $\beta_1 = 2(\sigma_a)^2$ and $\beta_2 = 2(\sigma_b)^2$. An event Ω of probability at least $1 - Ln^{-2}$ exists on which for every $m = (D_{m,1}, D_{m,2}) \in \mathcal{M}_n$,

$$\ell(s, \hat{s}_m) = \left(\frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,2}}{n} \right) (1 + \kappa_{m,0}) \quad (33)$$

$$\text{and} \quad P_n \gamma(\hat{s}_m) - P \gamma(s) = \left[\frac{\alpha}{D_{m,1}^2} + \frac{\alpha}{D_{m,2}^2} \right] (1 + \kappa_{m,1}) - \left[\frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,2}}{n} \right] (1 + \kappa_{m,2}) \quad , \quad (34)$$

where $(\kappa_{m,i})_{i=0,1,2, m \in \mathcal{M}_n}$ satisfy

$$\max \left\{ |\kappa_{m,i}| \mid s.t. \ i \in \{0, 1, 2\}, m \in \mathcal{M}_n, \text{ and } \min \{D_{m,1}, D_{m,2}\} \geq (\ln n)^6 \right\} \leq L_{(\mathbf{HTbm})} (\ln n)^{-1/2} \quad .$$

proof of Claim 3. Using the notation of Section 7.1,

$$\ell(s, \hat{s}_m) = \ell(s, s_m) + p_1(m) \quad \text{and} \quad P_n \gamma(\hat{s}_m) - P \gamma(s) = \ell(s, s_m) - p_2(m) + \bar{\delta}(m) \quad .$$

Let us first compute the expectation of each term. Recall that $\mathbb{E}[\bar{\delta}(m)] = 0$. The bias $\ell(s, s_m)$ is controlled thanks to Claim 2. By Proposition 4, $\mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)]$ mostly depend on

$$\begin{aligned} \sigma_\lambda^2 &= \mathbb{E} \left[(Y - s(X))^2 \mid X \in \lambda \right] \\ &= \mathbb{E} \left[(s(X) - s_m(X))^2 \mid X \in \lambda \right] + \mathbb{E} \left[(\sigma(X))^2 \mid X \in \lambda \right] \\ &= \frac{1}{\text{Leb}(\lambda)} \int_\lambda (s(x) - s_m(x))^2 dx + \frac{1}{\text{Leb}(\lambda)} \int_\lambda (\sigma(X))^2 dx \quad . \end{aligned}$$

Precisely,

$$\begin{aligned} \mathbb{E}[p_2(m)] &= \frac{1}{n} \sum_{\lambda \in \Lambda_m} \sigma_\lambda^2 \\ &= \frac{2D_{m,1}}{n} \int_0^{1/2} \left[(s(X) - s_m(x))^2 + (\sigma(x))^2 \right] dx + \frac{2D_{m,2}}{n} \int_{1/2}^1 \left[(s(X) - s_m(x))^2 + (\sigma(x))^2 \right] dx \\ &= \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,2}}{n} + R(m, n) \end{aligned}$$

$$\text{where} \quad 0 \leq R(m, n) = \frac{2}{n} \left(D_{m,1} \int_0^{1/2} (s(X) - s_m(x))^2 dx + D_{m,2} \int_{1/2}^1 (s(X) - s_m(x))^2 dx \right) \leq \frac{\ell(s, s_m)}{(\ln(n))^2} \quad ,$$

since $D_{m,i} \leq n/(2(\ln(n))^2)$. Similarly,

$$\begin{aligned} \mathbb{E}[p_1(m)] &= \left(\frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,2}}{n} \right) (1 + \delta_n) + R'(m, n) \\ \text{where} \quad 0 \leq R'(m, n) &\leq \frac{\ell(s, s_m)}{(\ln(n))^2} \quad \text{and} \quad |\delta_n| \leq L (\ln n)^{-1/2} \quad . \end{aligned}$$

It now remains to prove that $p_1(m) - \mathbb{E}[p_1(m)]$ and $\mathbb{E}[p_2(m)] - p_2(m) + \bar{\delta}(m)$ are close to zero on a large probability event. The condition on $\sigma(\cdot)$ imply that the last assumption of Proposition 5 holds since

$$D_m^{-1} \sum_{\lambda \in \Lambda_m} \mathbb{E}[\sigma(X)^2 \mid X \in \lambda] = \frac{2D_{m,1}(\sigma_a)^2 + 2D_{m,2}(\sigma_b)^2}{D_{m,1} + D_{m,2}} \geq 2 \min \{ (\sigma_a)^2, (\sigma_b)^2 \} =: Q > 0 \quad .$$

Let Ω_2 be the event on which, for every $m \in \mathcal{M}_n$, (20)–(22) hold with $\gamma = 4$ and $B_n = \ln(n)^2$, and (23) holds with $x = 4\ln(n)$ and $\eta = (\ln(n))^{-1}$. Since $\text{Card}(\mathcal{M}_n) \leq n^2$, Proposition 5 and Lemma 6 show that $\mathbb{P}(\Omega_2) \geq 1 - Ln^{-2}$. Therefore, the probability of $\Omega = \Omega_1 \cap \Omega_2$ is larger than $1 - K_1 n^{-2}$ for some absolute constant K_1 .

On Ω , for every $m \in \mathcal{M}_n$ such that $\min\{D_{m,1}, D_{m,2}\} \geq (\ln(n))^6$, we then have

$$\begin{aligned} |p_1(m) - \mathbb{E}[p_1(m)]| &\leq L_{(\mathbf{HThm})} (\ln n)^{-1} \mathbb{E}[p_2(m)] \\ |p_2(m) - \mathbb{E}[p_2(m)]| &\leq L_{(\mathbf{HThm})} (\ln n)^{-2} \mathbb{E}[p_2(m)] \\ \text{and} \quad |\bar{\delta}(m)| &\leq \frac{\ell(s, s_m)}{\ln n} + \frac{L_{(\mathbf{HThm})} (\ln n)^2}{n} \end{aligned}$$

as soon as $n \geq L$. Enlarging the constant K_1 so that $1 - K_1 n^{-2} \leq 0$ when n is too small yields the result. \square

Claim 4. On Ω ,

$$\inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\} \leq \frac{3}{2^{2/3} n^{2/3}} \left(\alpha_1^{1/3} \beta_1^{2/3} + \alpha_2^{1/3} \beta_2^{2/3} \right) \left(1 + L_{(\mathbf{HThm})} (\ln n)^{-1/2} \right). \quad (35)$$

proof of Claim 4. Let $m_* \in \mathcal{M}_n$ be any model such that

$$\left| D_{m_*,1} - \left(\frac{2\alpha_1 n}{\beta_1} \right)^{1/3} \right| \leq 1 \quad \text{and} \quad \left| D_{m_*,2} - \left(\frac{2\alpha_2 n}{\beta_2} \right)^{1/3} \right| \leq 1.$$

As soon as $n \geq L_{(\mathbf{HThm})}$, such an m_* exists and satisfies $\min\{D_{m_*,1}, D_{m_*,2}\} > (\ln n)^6$. The result follows from Claim 3. \square

Claim 5. For every $m \in \mathcal{M}_n$ such that $\min\{D_{m,1}, D_{m,2}\} \leq (\ln(n))^6$,

$$\ell(s, \hat{s}_m) \geq \frac{L_{\alpha_1, \alpha_2, \|s'\|_\infty, \|s''\|_\infty}}{(\ln(n))^{12}}.$$

In particular, Claims 4 and 5 show that for every $\mathbb{C}_1 > 0$, when $\min\{D_{m',1}, D_{m',2}\} \leq (\ln(n))^6$ and $n \geq L_{(\mathbf{HThm}), \mathbb{C}_1}$,

$$\ell(s, \hat{s}_{m'}) \geq \mathbb{C}_1 \inf_{m \in \mathcal{M}_n} \{\ell(s, \hat{s}_m)\}.$$

proof of Claim 5. First, note that $\ell(s, \hat{s}_m) \geq \ell(s, s_m)$, and by Claim 2, for every $m \in \mathcal{M}_n$,

$$\ell(s, s_m) \geq \alpha_1 D_{m,1}^{-2} \left(1 - L_{\|s'\|_\infty, \|s''\|_\infty} D_{m,1}^{-1} \right) + \alpha_2 D_{m,2}^{-2} \left(1 - L_{\|s'\|_\infty, \|s''\|_\infty} D_{m,2}^{-1} \right).$$

If $m \in \mathcal{M}_n$ satisfies $\min\{D_{m,1}, D_{m,2}\} \geq L_1 (\|s'\|_\infty, \|s''\|_\infty)$, then, the lower bound is larger than

$$\frac{\alpha_1}{2D_{m,1}^2} + \frac{\alpha_2}{2D_{m,2}^2} \geq \frac{\min\{\alpha_1, \alpha_2\}}{2 \min\{D_{m,1}, D_{m,2}\}^2}.$$

Now fix some $m \in \mathcal{M}_n$ such that $\min\{D_{m,1}, D_{m,2}\} \leq (\ln(n))^6$. Some $m' \in \mathcal{M}_n$ exists such that $S_m \subset S_{m'}$, $L_1 \leq D_{m',1} \leq 2 \max\{L_1, (\ln(n))^6\}$; indeed, either $m = m'$ satisfies the condition,

or m' can be obtained from m by doubling the number of bins in $[0, 1/2]$ (resp. $(1/2, 1]$) until the required condition is fulfilled. Then,

$$\ell(s, s_m) \geq \ell(s, s_{m'}) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2 \min\{D_{m,1}, D_{m,2}\}^2} \geq \frac{L_{\|s'\|_\infty, \|s''\|_\infty}}{(\ln(n))^{12}}.$$

□

Claim 6. Define for every $m \in \mathcal{M}_n$,

$$C_{m,1} := D_{m,1} \left(\frac{\beta_1}{2\alpha_1 n} \right)^{1/3} > 0 \quad \text{and} \quad C_{m,2} := D_{m,2} \left(\frac{\beta_2}{2\alpha_2 n} \right)^{1/3} > 0.$$

Let $\Delta \in (0, 1]$ and define

$$\eta_\Delta := \frac{\Delta^2}{17 \left(1 + \left(\frac{\beta_1}{\beta_2} \right)^{2/3} \right)}.$$

Then, on Ω , any $m \in \mathcal{M}_n$ such that

$$\ell(s, \widehat{s}_m) \leq (1 + \eta_\Delta) \inf_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m)\} \quad (36)$$

must satisfy

$$\max\{|C_{m,1} - 1|, |C_{m,2} - 1|\} \leq \Delta \quad (37)$$

as soon as $n \geq L_{(\mathbf{HThm}), \Delta}$.

proof of Claim 6. Assume that Ω holds and let $m \in \mathcal{M}_n$ be satisfying (36). From Claim 5, we know that $\min\{D_{m,1}, D_{m,2}\} > (\ln(n))^6$ for $n \geq L_{(\mathbf{HThm}), \Delta}$.

Define, for every $x > -1$, $f(x) = 2^{-2/3}(1+x)^{-2} + 2^{1/3}(1+x)$ and for every $x \geq 0$, $g(x) = \min\{1, (x-1)^2\}$. Then, (33) in Claim 3 and Lemma 8 below yield

$$\begin{aligned} \ell(s, \widehat{s}_m) &\geq \frac{1}{n^{2/3}} \left(\alpha_1^{1/3} \beta_1^{2/3} f(C_{m,1} - 1) + \alpha_2^{1/3} \beta_2^{2/3} f(C_{m,2} - 1) \right) \left(1 - L_{(\mathbf{HThm})} (\ln n)^{-1/2} \right) \\ &\geq \frac{3}{2^{2/3} n^{2/3}} \left(\alpha_1^{1/3} \beta_1^{2/3} + \alpha_2^{1/3} \beta_2^{2/3} \right) + \frac{3}{2^{14/3} n^{2/3}} \left(1 - L_{(\mathbf{HThm})} (\ln n)^{-1/2} \right) \left(\alpha_1^{1/3} \beta_1^{2/3} g(C_{m,1}) + \alpha_2^{1/3} \beta_2^{2/3} g(C_{m,2}) \right) \end{aligned}$$

Hence, (35) and (36) imply

$$16 \left(\alpha_1^{1/3} \beta_1^{2/3} + \alpha_2^{1/3} \beta_2^{2/3} \right) \frac{\eta_\Delta + L_{(\mathbf{HThm}), \Delta} (\ln n)^{-1/2}}{1 - L_{(\mathbf{HThm})} (\ln n)^{-1/2}} \geq \left(\alpha_1^{1/3} \beta_1^{2/3} g(C_{m,1}) + \alpha_2^{1/3} \beta_2^{2/3} g(C_{m,2}) \right).$$

In particular, when $n \geq L_{(\mathbf{HThm}), \Delta}$,

$$g(C_{m,1}) \leq 16\Delta^2/17 < 1 \quad \text{and} \quad g(C_{m,2}) \leq 16\Delta^2/17 < 1,$$

which implies (37). □

Lemma 8. Let $f : (-1, +\infty) \mapsto \mathbb{R}$ be defined by $f(x) = 2^{-2/3}(1+x)^{-2} + 2^{1/3}(1+x)$. Then, for every $x > -1$,

$$f(x) \geq 3 \times 2^{-2/3} + 3 \times 2^{-14/3} \min\{x^2, 1\}.$$

proof of Lemma 8. We apply the Taylor-Lagrange theorem to f (which is infinitely differentiable) at order two, between 0 and x . The result follows since $f(0) = 3 \times 2^{-2/3}$, $f'(0) = 0$ and $f''(t) = 6 \times 2^{-2/3} \times (1+t)^{-4} \geq 3 \times 2^{1/3-4}$ if $t \leq 1$. If $t > 1$, the result follows from the fact that $f' \geq 0$ on $[0, +\infty)$. \square

We now can conclude the proof of Theorem 2. Let us assume that on Ω , $m \in \mathcal{M}_n$ satisfies (36) for some $\Delta > 0$ to be chosen later. Without loss of generality, we can assume that $\sigma_a > \sigma_b$, hence $\beta_1 > \beta_2$. By Claim 6, we have

$$\left(\frac{2\alpha_1 n}{\beta_1}\right)^{1/3} (1 - \Delta) \leq D_{m,1} \leq \left(\frac{2\alpha_1 n}{\beta_1}\right)^{1/3} (1 + \Delta)$$

and

$$\left(\frac{2\alpha_2 n}{\beta_2}\right)^{1/3} (1 - \Delta) \leq D_{m,2} \leq \left(\frac{2\alpha_2 n}{\beta_2}\right)^{1/3} (1 + \Delta) .$$

Therefore,

$$D_{m,1} \leq \kappa D_{m,2} \quad \text{with} \quad \kappa = \frac{\alpha_1^{1/3} \beta_2^{1/3} (1 + \Delta)}{\alpha_2^{1/3} \beta_1^{1/3} (1 - \Delta)} .$$

Since $\beta_1 > \beta_2$, we can choose $\Delta = \Delta(\beta_1, \beta_2) > 0$ such that

$$\kappa \leq \kappa' = (\alpha_1/\alpha_2)^{1/3} (\beta_2/\beta_1)^{1/6} < \bar{\kappa} = (\alpha_1/\alpha_2)^{1/3} .$$

Note that $D_{m,1} \leq \kappa D_{m,2}$ is equivalent to $D_{m,1} \leq \kappa D_m / (1 + \kappa)$. Therefore, some $m' \in \mathcal{M}_n$ exists such that $D_m = D_{m'}$ and $-1 \leq D_{m',1} - D_m \bar{\kappa} / (1 + \bar{\kappa}) \leq 0$. Then, (34) implies that

$$\begin{aligned} P_n \gamma(\hat{s}_m) - P_n \gamma(\hat{s}_{m'}) &= \left[\frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} \right] (1 + \kappa_{m,1}) - \left[\frac{\alpha_1}{D_{m',1}^2} + \frac{\alpha_2}{D_{m',2}^2} \right] (1 + \kappa_{m',1}) \\ &\quad - \left[\frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,2}}{n} \right] (1 + \kappa_{m,2}) + \left[\frac{\beta_1 D_{m',1}}{n} + \frac{\beta_2 D_{m',2}}{n} \right] (1 + \kappa_{m',2}) \\ &\geq \left[\alpha_1 D_{m,1}^{-2} + \alpha_2 D_{m,2}^{-2} - \alpha_1 D_{m',1}^{-2} - \alpha_2 D_{m',2}^{-2} \right] \\ &\quad + \frac{\beta_1 (D_{m',1} - D_{m,1})}{n} + \frac{\beta_2 (D_{m',2} - D_{m,2})}{n} - L_{(\mathbf{HThm})} (\ln(n))^{-1/2} n^{-2/3} . \end{aligned}$$

Now, remark that the bias term is smaller for $S_{m'}$ than for S_m since $x \mapsto \alpha_1 x^{-2} + \alpha_2 (D_m - x)^{-2}$ is decreasing on $(0, D_m \bar{\kappa} / (1 + \bar{\kappa})]$. Therefore, using the definition of m' ,

$$\begin{aligned} P_n \gamma(\hat{s}_m) - P_n \gamma(\hat{s}_{m'}) &\geq \frac{\beta_1 (D_{m',1} - D_{m,1})}{n} + \frac{\beta_2 (D_{m',2} - D_{m,2})}{n} - L_{(\mathbf{HThm})} (\ln(n))^{-1/2} n^{-2/3} \\ &= \frac{(\beta_1 - \beta_2)(D_{m',1} - D_{m,1})}{n} - L_{(\mathbf{HThm})} (\ln(n))^{-1/2} n^{-2/3} \\ &\geq \frac{(\beta_1 - \beta_2) \left[D_m \left(\frac{1}{1+\kappa} - \frac{1}{1+\bar{\kappa}} \right) - 1 \right]}{n} - L_{(\mathbf{HThm})} (\ln(n))^{-1/2} n^{-2/3} \\ &\geq L_{(\mathbf{HThm})} n^{-2/3} - L_{(\mathbf{HThm})} (\ln(n))^{-1/2} n^{-2/3} > 0 \end{aligned}$$

as soon as $n \geq L_{(\mathbf{HThm})}$. Therefore, $m \notin \mathcal{M}_{\text{dim}}$, which concludes the proof of Theorem 2, with $\mathbb{C}_1 = 1 + \eta_\Delta$. \square

7.5 Proof of Proposition 2

Let $L_{(\mathbf{HPro2})} = L_{c_{\mathcal{M}}, \alpha_{\mathcal{M}}, c_{\text{rich}}, A, \sigma_{\min}, C_{\mathbf{b}}^+, C_{\mathbf{b}}^-, \beta_-, \beta_+, c_{\mathbf{r}, \ell}^X}$ denote a constant (varying from line to line) that may only depend on the constants appearing in the assumptions of Proposition 2 (except the constant K).

According to (19) in Proposition 4,

$$\mathbb{E}[p_2(m)] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \sigma_{\lambda}^2 \geq \frac{D_m \sigma_{\min}^2}{n} \quad \text{and} \quad \mathbb{E}[p_2(m)] \leq \frac{D_m \|\sigma\|_{\infty}^2}{n} + \frac{1}{n} \sum_{\lambda \in \Lambda_m} (\sigma_{\lambda}^d)^2.$$

Now, using $(\mathbf{Ar}_{\ell}^{\mathbf{X}})$ and (\mathbf{Ap}) ,

$$C_{\mathbf{b}}^+ D_m^{-\beta_+} \geq \ell(s, s_m) = \sum_{\lambda \in \Lambda_m} p_{\lambda} (\sigma_{\lambda}^d)^2 \geq \frac{c_{\mathbf{r}, \ell}^X}{D_m} \sum_{\lambda \in \Lambda_m} (\sigma_{\lambda}^d)^2$$

so that

$$\mathbb{E}[p_2(m)] \leq \frac{D_m}{n} \left(\|\sigma\|_{\infty}^2 + \frac{C_{\mathbf{b}}^+}{c_{\mathbf{r}, \ell}^X D_m^{\beta_+}} \right).$$

Therefore, for every $m \in \mathcal{M}_n$ such that $D_m \geq \ln(n)$,

$$c_1(K, n) \mathbb{E}[p_2(m)] \leq \text{pen}(m) \leq c_2 \mathbb{E}[p_2(m)]$$

with $c_2 = K/\sigma_{\min}^2$ and

$$c_1(K, n) = \frac{K}{\|\sigma\|_{\infty}^2 + \frac{C_{\mathbf{b}}^+}{c_{\mathbf{r}, \ell}^X (\ln(n))^{\beta_+}}} \geq \frac{K}{\|\sigma\|_{\infty}^2} \times \frac{1}{1 + \frac{C_{\mathbf{b}}^+}{c_{\mathbf{r}, \ell}^X (\ln(n))^{\beta_+} \|\sigma\|_{\infty}^2}} \geq \frac{K}{\|\sigma\|_{\infty}^2} \left(1 - (\ln(n))^{-\beta_+/2} \right) \geq \frac{1}{2} \left(1 + \frac{K}{\|\sigma\|_{\infty}^2} \right),$$

as soon as $n \geq L_{(\mathbf{HPro2}), K}$. Then, Theorem 5 in [9] shows that with probability at least $1 - L_{(\mathbf{HPro2})} n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left(\frac{1 + (\frac{K}{\sigma_{\min}^2} - 2)_+}{\min \left\{ 1, \frac{1}{2} \left(\frac{K}{\|\sigma\|_{\infty}^2} - 1 \right) \right\}} \right) \inf_{m \in \mathcal{M}_n} \{ \ell(s, \widehat{s}_m) \}$$

which concludes the proof. When $K \geq 2 \|\sigma\|_{\infty}^2$, the leading constant of the oracle inequality is smaller than

$$\left(\frac{1 + (\frac{K}{\sigma_{\min}^2} - 2)_+}{1 - (\ln(n))^{-\beta_+/2}} + (\ln(n))^{-1/5} \right)$$

which can be made as close as possible from $K \sigma_{\min}^{-2} - 1$ provided $n \geq L_{(\mathbf{HPro2}), K}$. \square

7.6 Proof of Proposition 3

Let $L_{(\mathbf{HPro3})} = L_{A, \sigma_{\min}, \alpha, R, \mu, c_{X, \text{Leb}}, \inf_{[0,1]} \sigma}$ denote a constant (varying from line to line) that may only depend on the constants appearing in the assumptions of Proposition 3 (except the constant K).

Since $\mathcal{M}_n = \mathcal{M}_n^{(\text{reg}, 1/2)}$ and the penalty can be written $\text{pen}(m) = \frac{KD_{m,1}}{n} + \frac{KD_{m,2}}{n}$, the model selection problem can actually be split into two separate model selection problems: one for the N data points for which $X_i \in [0, 1/2]$, the other for the $n - N$ data points for which $X_i \in (1/2, 1]$.

For proving Proposition 3, we can focus on the first problem only, that is, we are given N data points independent with distribution $\mathcal{L}((X, Y) \mid X \in [0, 1/2])$, where N is itself a random variable whose distribution is binomial with parameters n and $\mathbb{P}(X \in [0, 1/2]) = \mu$. The goal is to select a model \tilde{m} among the family $\tilde{\mathcal{M}}(n, N)$ of regular histograms on $[0, 1/2]$ with a number of bins between 1 and $n/(2\ln(n))^2$. Note that from Bernstein's inequality (see for instance Proposition 2.9 in [23]), we have with probability at least $1 - 2n^{-2}$ that

$$n\mu + 2\sqrt{\mu n \ln(n)} + \frac{2\ln(n)}{3} \geq N \geq n\mu - 2\sqrt{\mu n \ln(n)} - \frac{2\ln(n)}{3}.$$

In particular, on some event Ω_n of probability at least $1 - 2n^{-2}$,

$$\text{if } n \geq L_\mu, \quad \text{then} \quad \left| \frac{N}{n\mu} - 1 \right| \leq n^{-1/4}.$$

Now, on Ω_n , we apply Theorem 2 in [9]. First, let us check that the assumptions of Theorem 2 in [9] are satisfied: **(Ab)** and **(An)** are assumed in Proposition 3; the upper bound on the bias of the models holds because $s \in \mathcal{H}(\alpha, R)$; the uniform lower bound on $\mathbb{P}(X \in \lambda)$ holds because $\mathbb{P}(X \in [0, 1/2]) = \mu > 0$ and X has a lower bounded density w.r.t. $\text{Leb}([0, 1/2])$. Finally, we need an upper bound on $\text{pen}(\tilde{m}) = KD_{\tilde{m}}/N$: Using the proof of Proposition 2, we have

$$\forall \tilde{m} \in \tilde{\mathcal{M}}(n, N), \quad \frac{\text{pen}(\tilde{m})}{\mathbb{E}[p_2(\tilde{m}, N)]} \leq \frac{KD_{\tilde{m}}N}{ND_{\tilde{m}} \inf_{t \in [0, 1/2]} \{\sigma(t)^2\}} = \frac{K}{\inf_{t \in [0, 1/2]} \{\sigma(t)^2\}} < 1.$$

So, Theorem 2 in [9] shows that $D_{\tilde{m},1} \geq L_{(\text{HPro3}),K} N (\ln(N))^{-2} \geq L_{(\text{HPro3}),K} n (\ln(n))^{-1}$ with probability at least $1 - L_{(\text{HPro3}),K} N^{-2} = 1 - L_{(\text{HPro3}),K} n^{-2}$. The lower bound (13) on the risk also follows from Theorem 2 in [9] and its proof. \square

Acknowledgments

The author would like to thank gratefully Pascal Massart for several fruitful discussions, and Francis Bach for pointing out an idea that led to improve Theorem 2. The author acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01.

References

- [1] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.
- [2] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.
- [3] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

- [4] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. <http://tel.archives-ouvertes.fr/tel-00198803/>.
- [5] Sylvain Arlot. *V-fold cross-validation improved: V-fold penalization*, February 2008. arXiv:0802.0566v2.
- [6] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624 (electronic), 2009.
- [7] Sylvain Arlot. Technical appendix to “Model selection by resampling penalization”, 2009. Appendix to Electronic Journal of Statistics, 3, (2009), 557-624 (electronic). hal-00262478.
- [8] Sylvain Arlot and Alain Celisse. Segmentation in the mean of heteroscedastic data via cross-validation, April 2009. arXiv:0902.3977v2.
- [9] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.
- [10] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.
- [11] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.
- [12] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
- [13] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- [14] Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression: X -fixed prediction error. *J. Amer. Statist. Assoc.*, 87(419):738–754, 1992.
- [15] Bradley Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
- [16] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.
- [17] Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642, 2004. With comments and a rejoinder by the author.
- [18] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.
- [19] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.*, 2:1345–1372, 2008.
- [20] Matthieu Lerasle. Optimal model selection in density estimation. Preprint, 2009.
- [21] Ker-Chau Li. Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.
- [22] Colin L. Mallows. Some comments on C_p . *Technometrics*, 15:661–675, 1973.

- [23] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- [24] Boris T. Polyak and A. B. Tsybakov. Asymptotic optimality of the C_p -test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, 35(2):305–317, 1990.
- [25] Marie Sauvé. Histogram selection in non gaussian regression. *ESAIM: Probability and Statistics*, 13:70–86, March 2009.
- [26] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.
- [27] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.
- [28] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

A Appendix

Table 4: Accuracy indices C_{or} for each procedure in the four experiments $\pm \varepsilon_{C_{\text{or}}, N}$. In each column, the most accurate procedures (taking the uncertainty $\varepsilon_{C_{\text{or}}, N}$ into account) are bolded.

Experiment	X1-005	S0-1	XS1-05	X1-005 μ 02
Mal	8.961 \pm 0.055	6.056 \pm 0.033	3.424 \pm 0.020	4.693 \pm 0.019
Mal \times 1.25	7.279 \pm 0.057	5.086 \pm 0.034	2.347 \pm 0.015	4.692 \pm 0.019
Mal \times 2	4.101 \pm 0.036	3.242 \pm 0.022	1.452 \pm 0.005	4.562 \pm 0.021
Mal \times 3	3.074 \pm 0.019	2.675 \pm 0.013	1.338 \pm 0.003	4.033 \pm 0.023
Mal \times 4	2.862 \pm 0.015	3.214 \pm 0.017	1.891 \pm 0.014	3.420 \pm 0.022
Mal $_{\infty}$	4.110 \pm 0.036	3.242 \pm 0.022	1.664 \pm 0.008	3.122 \pm 0.020
Mal $_{\infty} \times 1.25$	3.371 \pm 0.024	2.794 \pm 0.015	1.452 \pm 0.005	3.334 \pm 0.015
Mal $_{\infty} \times 2$	2.862 \pm 0.015	3.214 \pm 0.017	1.358 \pm 0.004	3.370 \pm 0.009
Mal $_{\infty} \times 3$	3.033 \pm 0.019	5.035 \pm 0.015	3.493 \pm 0.019	3.430 \pm 0.007
Mal $_{\infty} \times 4$	3.549 \pm 0.025	5.810 \pm 0.006	5.020 \pm 0.009	3.493 \pm 0.006
HO	3.598 \pm 0.036	2.707 \pm 0.021	1.848 \pm 0.008	2.398 \pm 0.018
2FCV	3.104 \pm 0.032	2.458 \pm 0.019	1.767 \pm 0.007	2.289 \pm 0.016
5FCV	3.176 \pm 0.035	2.538 \pm 0.021	1.749 \pm 0.008	2.332 \pm 0.018
10FCV	3.291 \pm 0.037	2.559 \pm 0.022	1.738 \pm 0.008	2.369 \pm 0.018
penHO	5.070 \pm 0.045	3.492 \pm 0.027	2.529 \pm 0.014	2.798 \pm 0.020
penHO \times 1.25	4.393 \pm 0.041	3.072 \pm 0.024	2.152 \pm 0.012	2.626 \pm 0.019
penHO \times 2	3.595 \pm 0.034	2.659 \pm 0.020	1.853 \pm 0.008	2.751 \pm 0.034
penHO \times 3	3.516 \pm 0.032	2.558 \pm 0.018	1.972 \pm 0.009	3.634 \pm 0.055
penHO \times 4	3.729 \pm 0.033	2.573 \pm 0.018	2.166 \pm 0.011	4.663 \pm 0.070
pen2F	4.530 \pm 0.043	3.229 \pm 0.025	2.325 \pm 0.013	2.729 \pm 0.019
pen2F \times 1.25	3.649 \pm 0.037	2.769 \pm 0.022	1.945 \pm 0.010	2.451 \pm 0.018
pen2F \times 2	2.619 \pm 0.028	2.270 \pm 0.017	1.619 \pm 0.005	2.062 \pm 0.014
pen2F \times 3	2.273 \pm 0.023	2.222 \pm 0.015	1.539 \pm 0.005	1.932 \pm 0.013
pen2F \times 4	2.275 \pm 0.022	2.381 \pm 0.016	1.586 \pm 0.007	1.907 \pm 0.014
pen5F	3.779 \pm 0.041	2.857 \pm 0.024	1.925 \pm 0.010	2.540 \pm 0.019
pen5F \times 1.25	2.794 \pm 0.031	2.331 \pm 0.018	1.646 \pm 0.007	2.193 \pm 0.016
pen5F \times 2	2.051 \pm 0.019	1.995 \pm 0.012	1.457 \pm 0.004	1.880 \pm 0.011
pen5F \times 3	1.777 \pm 0.013	2.119 \pm 0.011	1.388 \pm 0.003	1.860 \pm 0.009
pen5F \times 4	1.838 \pm 0.015	2.384 \pm 0.013	1.366 \pm 0.003	1.887 \pm 0.008
pen10F	3.599 \pm 0.040	2.726 \pm 0.024	1.810 \pm 0.009	2.451 \pm 0.019
pen10F \times 1.25	2.726 \pm 0.031	2.215 \pm 0.018	1.594 \pm 0.006	2.125 \pm 0.016
pen10F \times 2	1.893 \pm 0.016	1.944 \pm 0.012	1.451 \pm 0.004	1.854 \pm 0.010
pen10F \times 3	1.709 \pm 0.011	2.132 \pm 0.011	1.358 \pm 0.003	1.879 \pm 0.008
pen10F \times 4	1.706 \pm 0.011	2.389 \pm 0.011	1.327 \pm 0.002	1.943 \pm 0.007
penLoo	3.171 \pm 0.034	2.499 \pm 0.021	1.731 \pm 0.008	2.395 \pm 0.019
penLoo \times 1.25	2.529 \pm 0.027	2.118 \pm 0.016	1.548 \pm 0.006	2.065 \pm 0.015
penLoo \times 2	1.870 \pm 0.014	1.954 \pm 0.012	1.401 \pm 0.003	1.879 \pm 0.009
penLoo \times 3	1.701 \pm 0.010	2.183 \pm 0.011	1.378 \pm 0.003	1.931 \pm 0.007
penLoo \times 4	1.679 \pm 0.010	2.457 \pm 0.011	1.308 \pm 0.002	2.002 \pm 0.006
Epenid	2.805 \pm 0.029	2.291 \pm 0.019	1.702 \pm 0.008	2.333 \pm 0.019
Epenid \times 1.25	2.304 \pm 0.023	1.943 \pm 0.014	1.513 \pm 0.005	2.035 \pm 0.015
Epenid \times 2	1.780 \pm 0.012	1.897 \pm 0.011	1.371 \pm 0.003	1.868 \pm 0.009
Epenid \times 3	1.687 \pm 0.009	2.161 \pm 0.011	1.312 \pm 0.002	1.938 \pm 0.007
Epenid \times 4	1.646 \pm 0.009	2.448 \pm 0.010	1.299 \pm 0.002	2.005 \pm 0.006

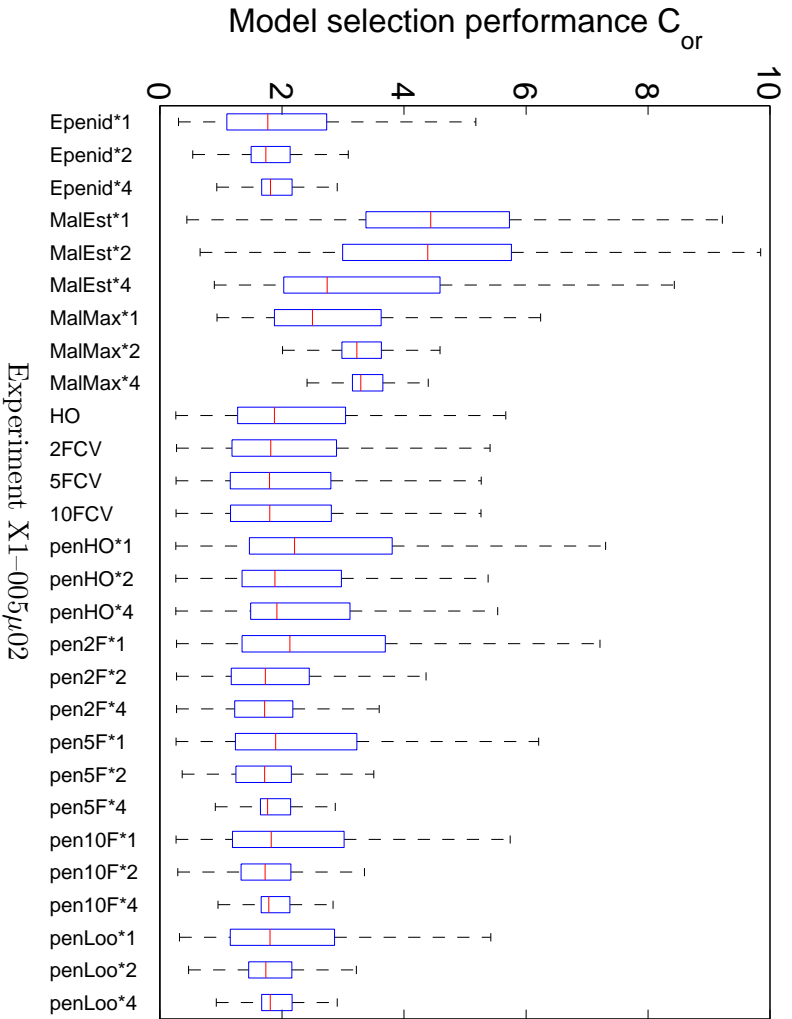
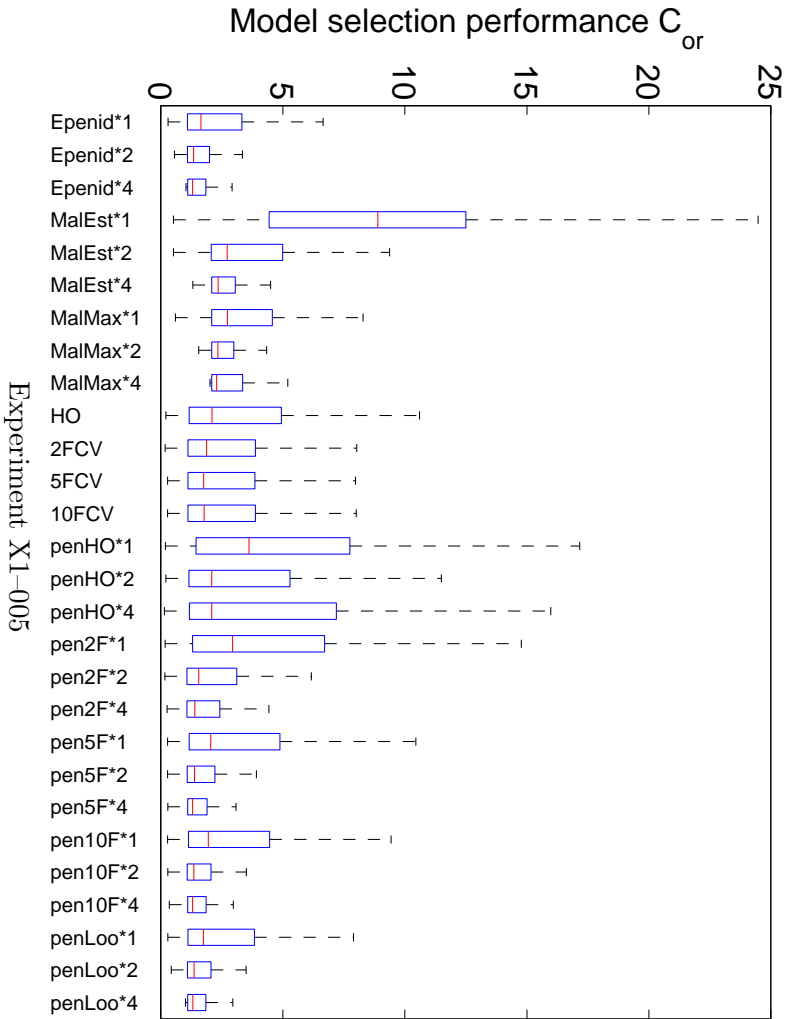


Figure 9: Box plot of $\ell(s, \hat{s}_m)$ divided by the estimated value of $\mathbb{E}[\ell(s, \hat{s}_m^*)]$ for various algorithms in experiments X1-005 and X1-005, μ_{02} .

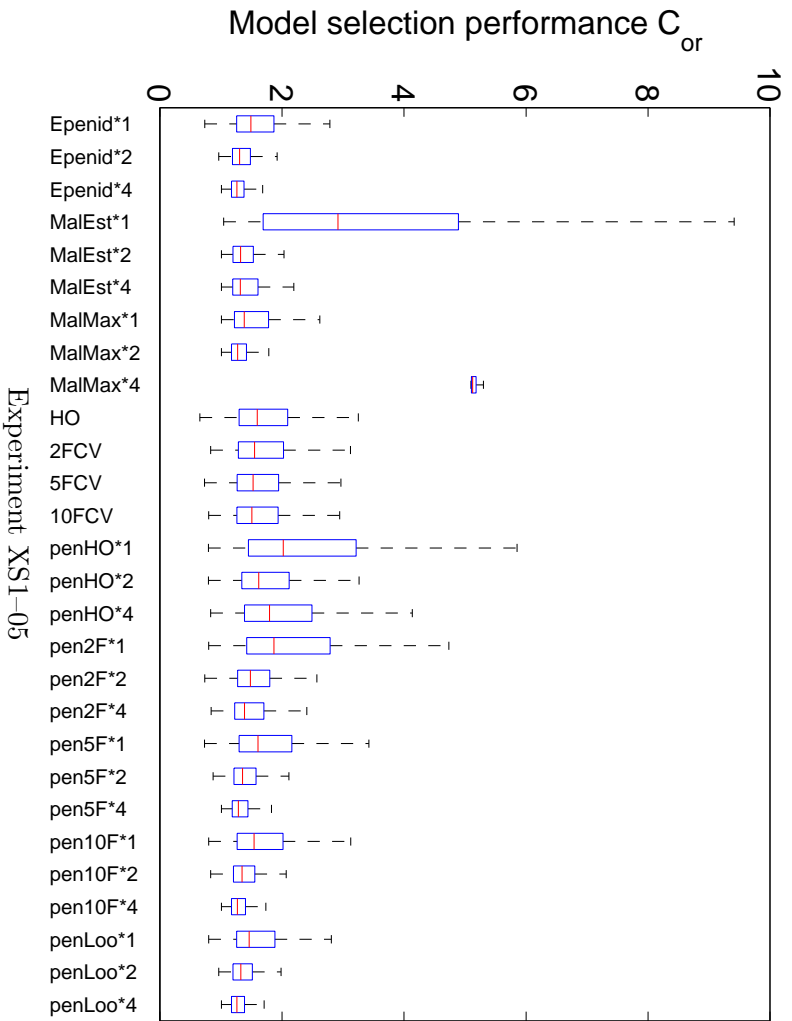
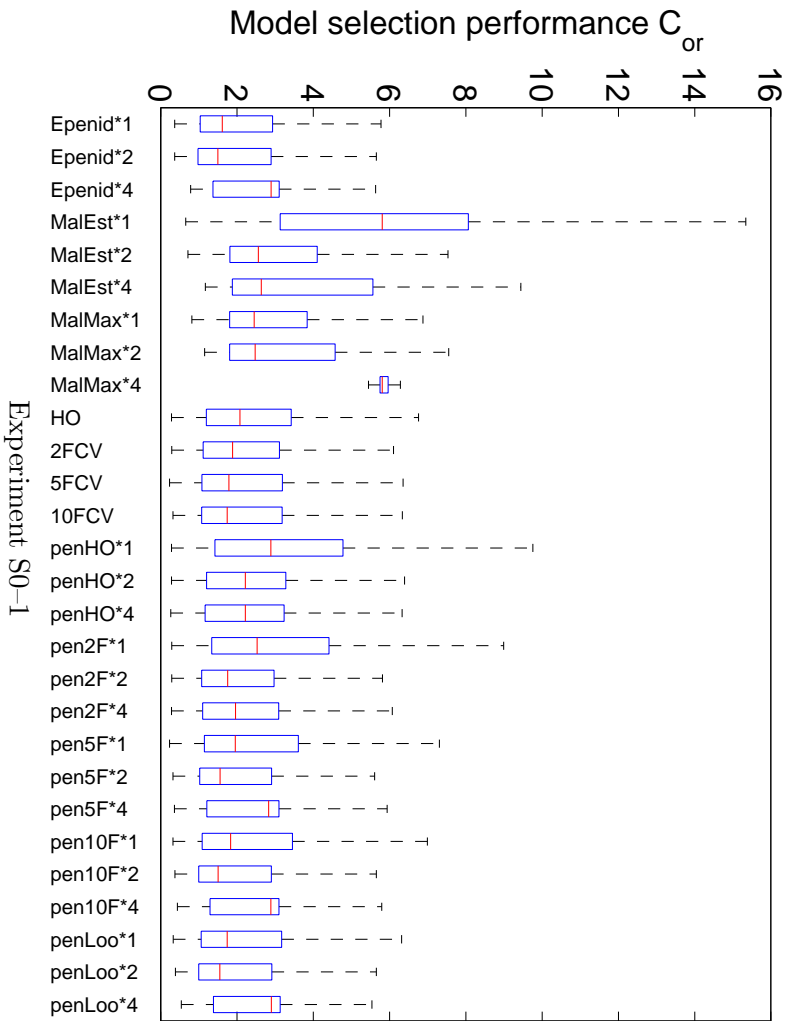


Figure 10: Box plot of $\ell(s, \hat{s}_m)$ divided by the estimated value of $\mathbb{E}[\ell(s, \hat{s}_m^*)]$ for various algorithms in experiments S0-1 and XS1-05.

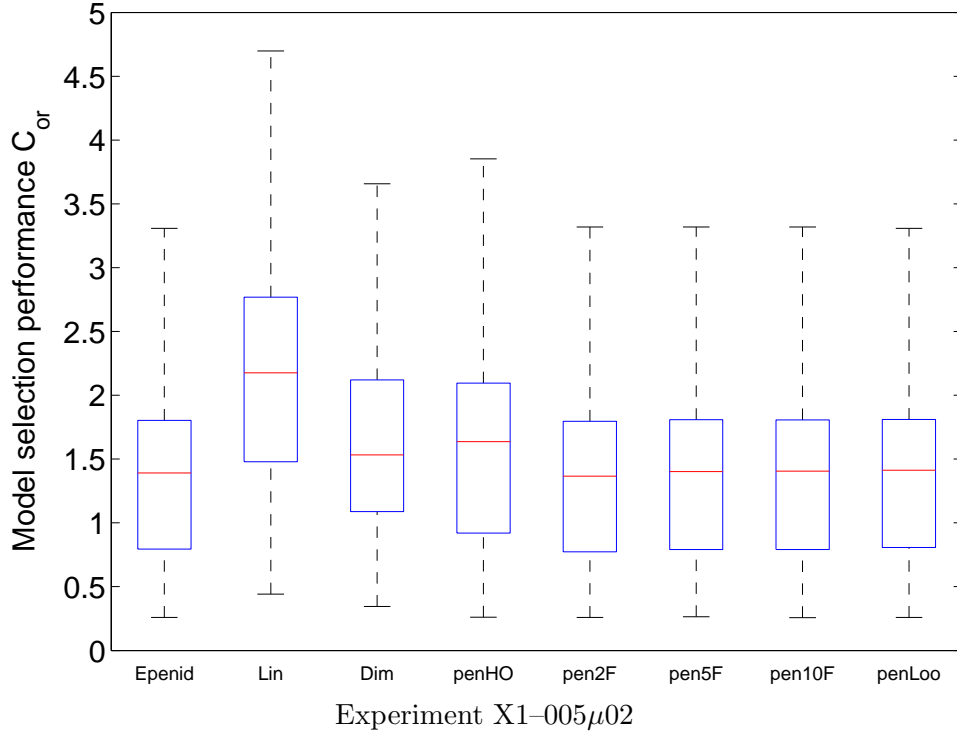
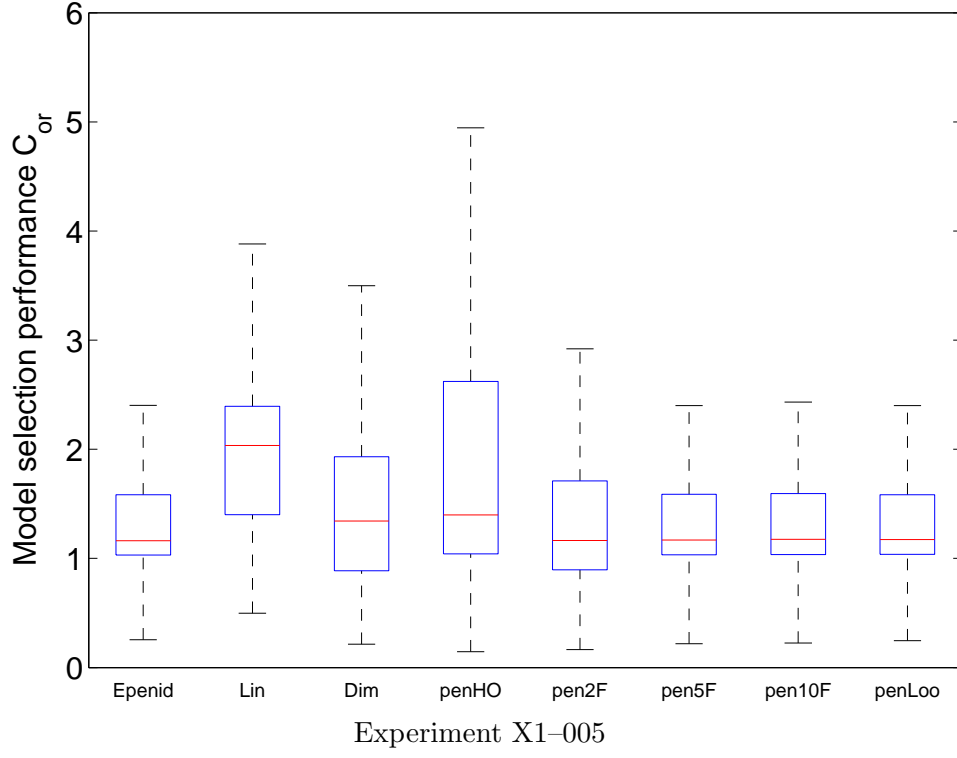


Figure 11: Box plot of $\ell(s, \hat{s}_{\hat{m}})$ divided by the estimated value of $\mathbb{E}[\ell(s, \hat{s}_{m^*})]$ for algorithms $\text{Id}\star$ (that is, penalties with the optimal data-driven overpenalization factor) in experiments X1-005 and X1-005μ02.

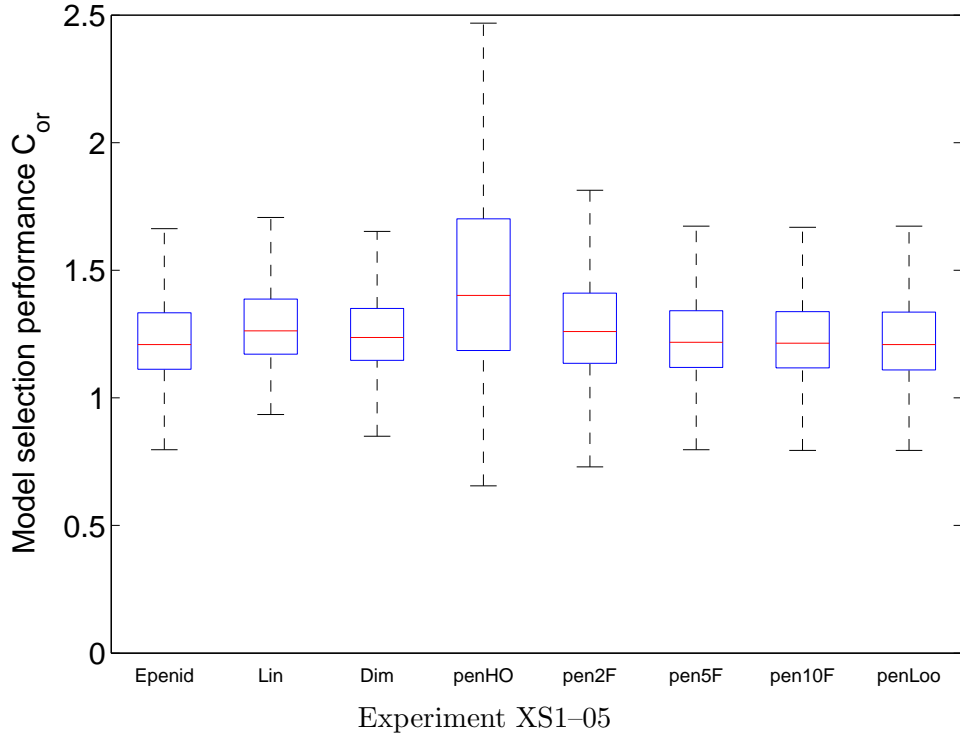
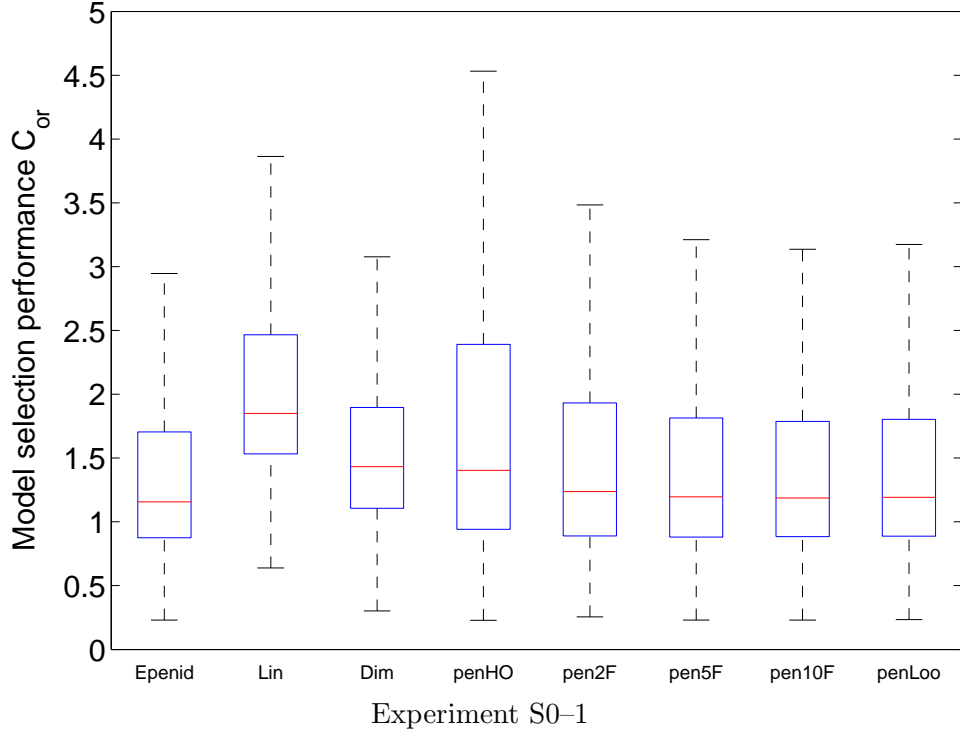


Figure 12: Box plot of $\ell(s, \hat{s}_{\hat{m}})$ divided by the estimated value of $\mathbb{E}[\ell(s, \hat{s}_{m^*})]$ for algorithms Id \star (that is, penalties with the optimal data-driven overpenalization factor) in experiments S0-1 and XS1-05.

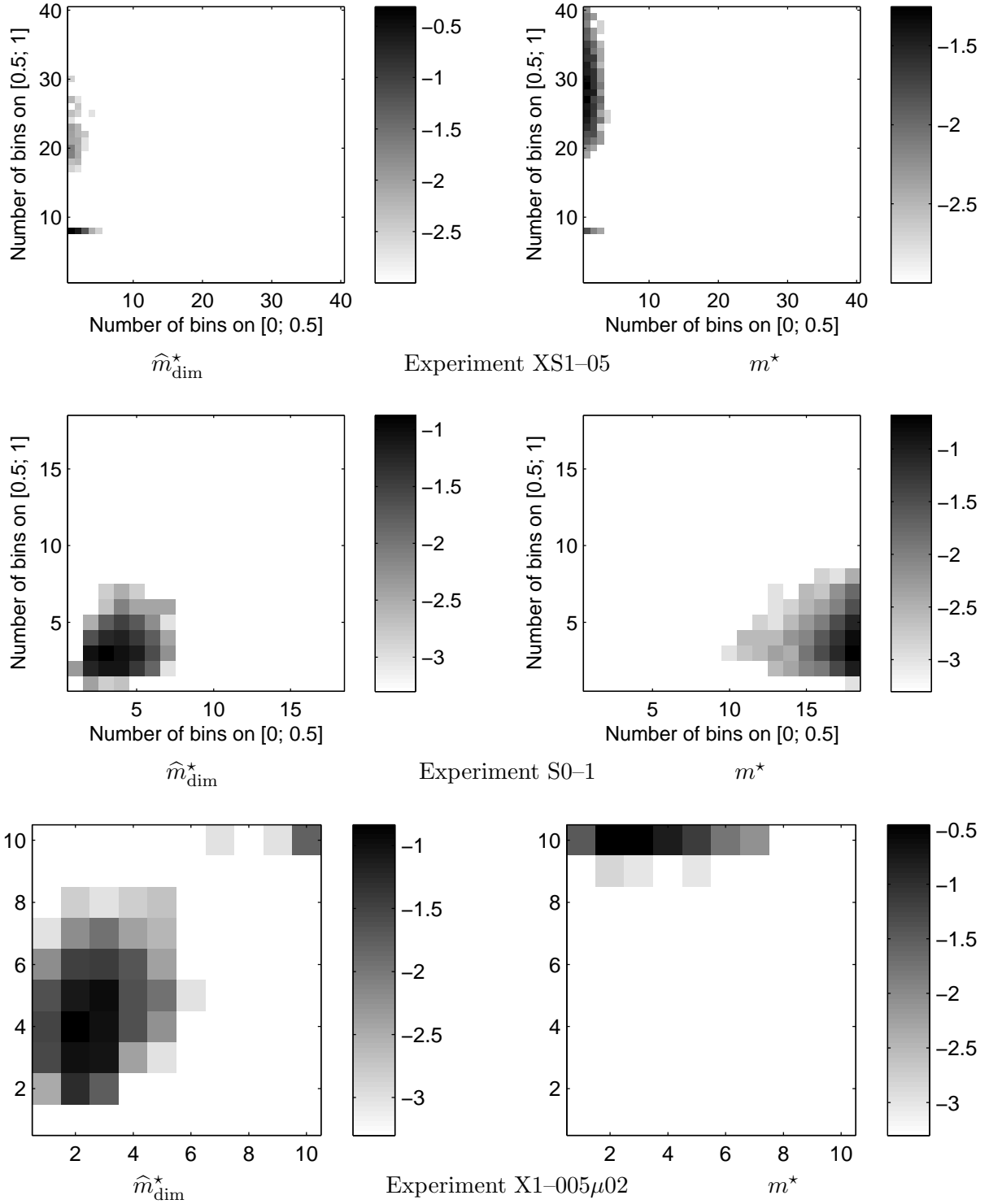


Figure 13: Same as Figure 3 for the three other experiments. Left: $\log_{10} \mathbb{P}(m = \hat{m}_{\text{dim}}^*)$ represented in \mathbb{R}^2 using $(D_{m,1}, D_{m,2})$ as coordinates, where \hat{m}_{dim}^* is defined by (9); $N = 10\,000$ samples have been simulated for estimating the probabilities. Right: $\log_{10} \mathbb{P}(m = m^*)$ using the same representation and the same samples. The distributions of m^* and \hat{m}_{dim}^* are almost disjoint.